Introduction to Statistical Signal Processing

Digital Signal Processing

- Digital ≡ sampled, discrete-time, quantized
- Signal ≡ waveform, sequnce of measurements or observations
- Processing ≡ analyze, modify, filter, synthesize

Examples of Digital Signals

- sampled speech waveform
- "pixelized" image
- Dow-Jones Index

DSP Applications

- Filtering (noise reduction)
- Pattern recognition (speech, faces, fingerprints)
- Compression

A Major Difficulty

In many (perhaps most) DSP applications we don't have complete or perfect knowledge of the signals we wish to process. We are faced with many **unknowns** and **uncertainties**.

Examples

- noisy measurements
- unknown signal parameters
- noisy system or environmental conditions
- natural variability in the signals encountered

Functional Magnetic Resonance Imaging
[missing_resource: FMRI.png]

Challenges are measurement noise and intrinsic uncertainties in signal behavior.

How can we design signal processing algorithms in the face of such uncertainty?

Can we model the uncertainty and incorporate this model into the design process?

Statistical signal processing is the study of these questions.

Modeling Uncertainty

The most widely accepted and commonly used approach to modeling uncertainty is **Probability Theory** (although other alternatives exist such as Fuzzy Logic).

Probability Theory models uncertainty by specifying the **chance** of observing certain signals.

Alternatively, one can view probability as specifying the degree to which we **believe** a signal reflects the true **state of nature**.

Examples of Probabilistic Models

- errors in a measurement (due to an imprecise measuring device) modeled as realizations of a Gaussian random variable.
- uncertainty in the phase of a sinusoidal signal modeled as a uniform random variable on $[0, 2\pi)$.
- uncertainty in the number of photons stiking a CCD per unit time modeled as a Poisson random variable.

Statistical Inference

A **statistic** is a function of observed data.

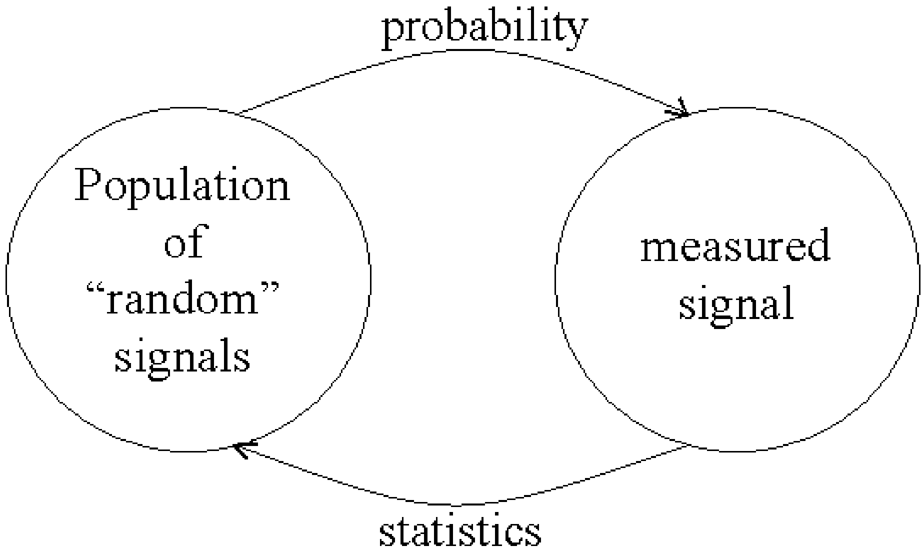
Example:
Suppose we observe N scalar values $x_1, \ldots, x_N$. The following are statistics:

- $\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$ (sample mean)
- $x_1, \ldots, x_N$ (the data itself)
- $\min \{x_1, \ldots, x_N\}$ (order statistic)
- $({x_1}^2 - x_2 \sin(x_3), e^{-(x_1 x_3)})$

A statistic **cannot** depend on **unknown parameters**.

Probability is used to model uncertainty.

Statistics are used to draw conclusions about probability models.



Probability models our uncertainty about signals we **may** observe.

Statistics reasons from the measured signal to the population of possible signals.

Statistical Signal Processing

- **Step 1** Postulate a probability model (or models) that reasonably capture the uncertainties at hand
- **Step 2** Collect data
- **Step 3** Formulate statistics that allow us to interpret or understand our probability model(s)

In this class

The two major kinds of problems that we will study are **detection** and **estimation**. Most SSP problems fall under one of these two headings.

Detection Theory

Given two (or more) probability models, which on best explains the signal?

Examples

1. Decode wireless comm signal into string of 0's and 1's
2. Pattern recognition
    - voice recognition
    - face recognition
    - handwritten character recognition
3. Anomaly detection
    - radar, sonar
    - irregular, heartbeat
    - gamma-ray burst in deep space

Estimation Theory

If our probability model has free parameters, what are the best parameter settings to describe the signal we've observed?

Examples

1. Noise reduction
2. Determine parameters of a sinusoid (phase, amplitude, frequency)
3. Adaptive filtering
    - track trajectories of space-craft
    - automatic control systems
    - channel equalization
4. Determine location of a submarine (sonar)
5. Seismology: estimate depth below ground of an oil deposit

Example:

Detection Example

Suppose we observe N tosses of an unfair coin. We would like to decide which side the coin favors, heads or tails.

- **Step 1** Assume each toss is a realization of a Bernoulli random variable.

  $$\mathrm{Pr}[\text{Heads}] = p = 1 - \mathrm{Pr}[\text{Tails}]$$

  Must decide $p = \frac{1}{4}$ vs. $p = \frac{3}{4}$.
- **Step 2** Collect data $x_1, \ldots, x_N$

  $$x_i = 1 \equiv \text{Heads}$$

  $$x_i = 0 \equiv \text{Tails}$$

- **Step 3** Formulate a useful statistic

  $$k = \sum_{n=1}^{N} x_n$$

  If $k < \frac{N}{2}$, guess $p = \frac{1}{4}$. If $k > \frac{N}{2}$, guess $p = \frac{3}{4}$.

Example:

Estimation Example

Suppose we take N measurements of a DC voltage A with a noisy voltmeter. We would like to estimate A.

- **Step 1** Assume a Gaussian noise model

  $$x_n = A + w_n$$

  where $w_n \sim \quad (0, 1)$.
- **Step 2** Gather data $x_1, \ldots, x_N$
- **Step 3** Compute the sample mean,

$$\widehat{A} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

and use this as an estimate.

In these examples ([link] and [link]), we solved detection and estimation problems using intuition and heuristics (in Step 3).

This course will focus on developing principled and mathematically rigorous approaches to detection and estimation, using the theoretical framework of probability and statistics.

Summary

- DSP ≡ processing signals with computer algorithms.
- SSP ≡ statistical DSP ≡ processing in the presence of uncertainties and unknowns.

Review of Linear Algebra

Vector spaces are the principal object of study in linear algebra. A vector space is always defined with respect to a field of scalars.

Fields

A field is a set F equipped with two operations, addition and mulitplication, and containing two special members 0 and 1 ($0 \neq 1$), such that for all $\{a, b, c\} \in F$

1. 1. $(a + b) \in F$
   2. $a + b = b + a$
   3. $(a + b) + c = a + (b + c)$
   4. $a + 0 = a$
   5. there exists $-a$ such that $a + -a = 0$

2. 1. $ab \in F$
   2. $ab = ba$
   3. $(ab)c = a\,(bc)$
   4. $a \cdot 1 = a$
   5. there exists a^{-1} such that $aa^{-1} = 1$

3. $a\,(b + c) = ab + ac$

More concisely

1. F is an abelian group under addition
2. F is an abelian group under multiplication
3. multiplication distributes over addition

Examples

$\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$

Vector Spaces

Let F be a field, and V a set. We say V **is a vector space over** F if there exist two operations, defined for all $a \in F$, $\boldsymbol{u} \in V$ and $\boldsymbol{v} \in V$:

- vector addition: $(\boldsymbol{u}, \boldsymbol{v})$ → $(\boldsymbol{u} + \boldsymbol{v}) \in V$
- scalar multiplication: $(a, \boldsymbol{v})$ → $a\boldsymbol{v} \in V$

and if there exists an element denoted $\mathbf{0} \in V$, such that the following hold for all $a \in F$, $b \in F$, and $\boldsymbol{u} \in V$, $\boldsymbol{v} \in V$, and $\boldsymbol{w} \in V$

1.
    1. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$
    2. $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$
    3. $\boldsymbol{u} + \mathbf{0} = \boldsymbol{u}$
    4. there exists $-\boldsymbol{u}$ such that $\boldsymbol{u} + -\boldsymbol{u} = \mathbf{0}$

2.
    1. $a\left(\boldsymbol{u} + \boldsymbol{v}\right) = a\boldsymbol{u} + a\boldsymbol{v}$
    2. $(a + b)\boldsymbol{u} = a\boldsymbol{u} + b\boldsymbol{u}$
    3. $(ab)\boldsymbol{u} = a\left(b\boldsymbol{u}\right)$
    4. $1 \cdot \boldsymbol{u} = \boldsymbol{u}$

More concisely,

1. V is an abelian group under plus
2. Natural properties of scalar multiplication

Examples

- $\mathbb{R}^N$ is a vector space over $\mathbb{R}$
- $\mathbb{C}^N$ is a vector space over $\mathbb{C}$
- $\mathbb{C}^N$ is a vector space over $\mathbb{R}$
- $\mathbb{R}^N$ is **not** a vector space over $\mathbb{C}$

The elements of V are called **vectors**.

Euclidean Space

Throughout this course we will think of a signal as a vector

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \dots & x_N \end{pmatrix}^T$$

The samples $\{x_i\}$ could be samples from a finite duration, continuous time signal, for example.

A signal will belong to one of two vector spaces:

Real Euclidean space

$\boldsymbol{x} \in \mathbb{R}^N$ (over $\mathbb{R}$)

Complex Euclidean space

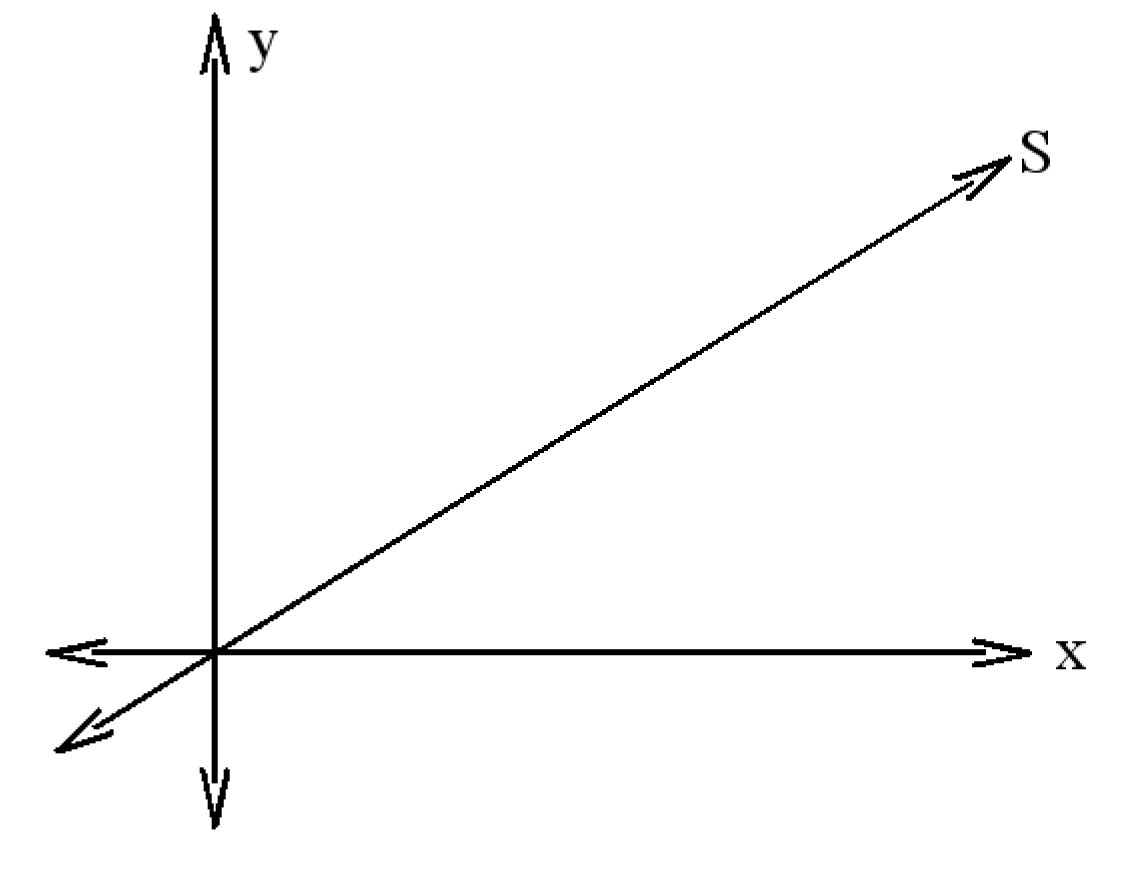
$\boldsymbol{x} \in \mathbb{C}^N$ (over $\mathbb{C}$)

Subspaces

Let V be a vector space over F.

A subset $S \subseteq V$ is called a **subspace** of V if S is a vector space over F in its own right.

Example:
$V = \mathbb{R}^2$, $F = \mathbb{R}$, $S =$ any line though the origin.



S is any line through the origin.

Are there other subspaces?

$S \subseteq V$ is a subspace if and only if for all $a \in F$ and $b \in F$ and for all $\boldsymbol{s} \in S$ and $\boldsymbol{t} \in S$, $(a\boldsymbol{s} + b\boldsymbol{t}) \in S$

Linear Independence

Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in V$.

We say that these vectors are **linearly dependent** if there exist scalars $a_1, \ldots, a_k \in F$ such that

Equation:

$$\sum_{i=1}^{k} a_i u_i = \mathbf{0}$$

and at least one $a_i \neq 0$.

If [link] only holds for the case $a_1 = \ldots = a_k = 0$, we say that the vectors are **linearly independent**.

Example:

$$1\begin{pmatrix}1\\-1\\2\end{pmatrix} - 2\begin{pmatrix}-2\\3\\0\end{pmatrix} + 1\begin{pmatrix}-5\\7\\-2\end{pmatrix} = \mathbf{0}$$

so these vectors are linearly dependent in $\mathbb{R}^3$.

Spanning Sets

Consider the subset $S = \{v_1, v_2, \ldots, v_k\}$. Define the **span** of S

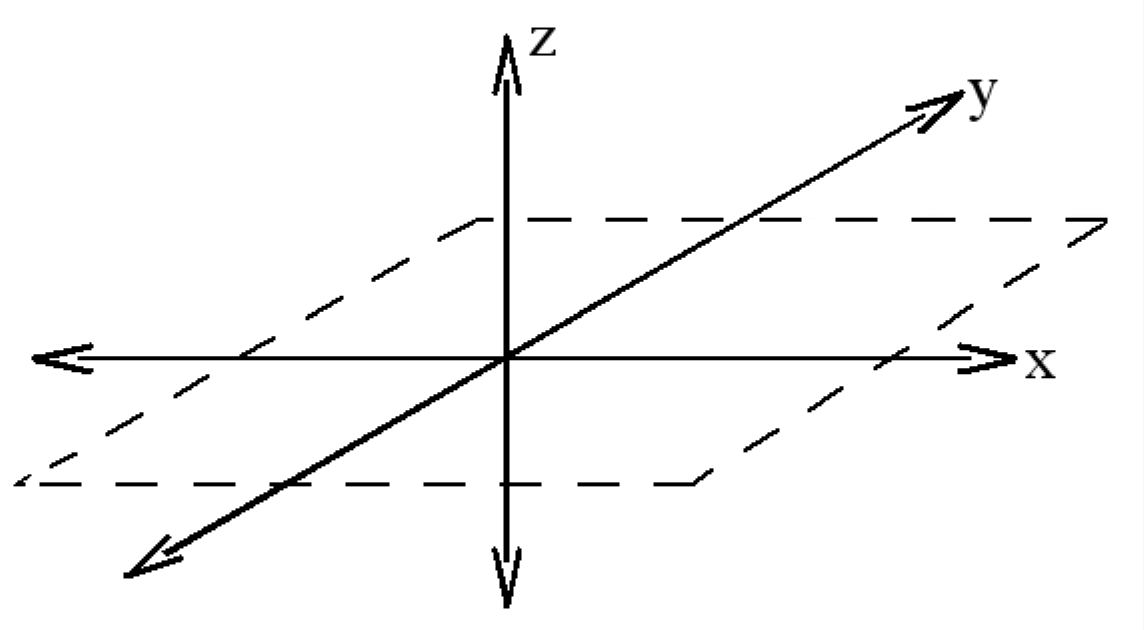
$$< S > \equiv \text{span}\,(S) \equiv \left\{ \sum_{i=1}^{k} a_i v_i \,\middle|\, a_i \in F \right\}$$

Fact: $< S >$ is a subspace of V.

Example:

$V = \mathbb{R}^3$, $F = \mathbb{R}$, $S = \{v_1, v_2\}$, $v_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $v_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \Rightarrow$

$< S >=$ xy-plane.



$< S >$ is the xy-plane.

Aside

If S is infinite, the notions of linear independence and span are easily generalized:

We say S is linearly independent if, for every finite collection $u_1, \ldots, u_k \in S$, (k arbitrary) we have

$$\left(\sum_{i=1}^{k} a_i u_i = \mathbf{0} \right) \Rightarrow \forall i : (a_i = 0)$$

The span of S is

$$< S >= \left\{ \sum_{i=1}^{k} a_i u_i \middle| a_i \in F \ \land \ u_i \in S \ \land \ (k < \infty) \right\}$$

Note: In both definitions, we only consider **finite** sums.

Bases

A set $B \subseteq V$ is called a **basis** for V over F if and only if

1. B is linearly independent
2. $< B >= V$

Bases are of fundamental importance in signal processing. They allow us to decompose a signal into building blocks (basis vectors) that are often more easily understood.

Example:
V = (real or complex) Euclidean space, $\mathbb{R}^N$ or $\mathbb{C}^N$.

$$B = \{e_1, \dots, e_N\} \equiv \text{standard basis}$$

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

where the 1 is in the i^{th} position.

Example:
$V = \mathbb{C}^N$ over $\mathbb{C}$.

$$B = \{u_1, \ldots, u_N\}$$

which is the DFT basis.

$$u_k = \begin{pmatrix} 1 \\ e^{-\left(i2\pi\frac{k}{N}\right)} \\ \vdots \\ e^{-\left(i2\pi\frac{k}{N}(N-1)\right)} \end{pmatrix}$$

where $i = \sqrt{-1}$.

Key Fact

If B is a basis for V, then every $\boldsymbol{v} \in V$ can be written uniquely (up to order of terms) in the form

$$\boldsymbol{v} = \sum_{i=1}^{N} a_i v_i$$

where $a_i \in F$ and $v_i \in B$.

Other Facts

- If S is a linearly independent set, then S can be extended to a basis.
- If $< S >= V$, then S contains a basis.

Dimension

Let V be a vector space with basis B. The dimension of V, denoted $\dim(V)$, is the cardinality of B.

Every vector space has a basis.

Every basis for a vector space has the same cardinality.

$\Rightarrow \dim(V)$ is **well-defined**.

If $\dim(V) < \infty$, we say V is **finite dimensional**.

Examples

vector space	**field of scalars**	**dimension**
$\mathbb{R}^N$	$\mathbb{R}$	
$\mathbb{C}^N$	$\mathbb{C}$	
$\mathbb{C}^N$	$\mathbb{R}$	

Every subspace is a vector space, and therefore has its own dimension.

Example:
Suppose $(S = \{u_1, \ldots, u_k\}) \subseteq V$ is a linearly independent set. Then

$$\dim(< S >) =$$

Facts

- If S is a subspace of V, then $\dim(S) \leq \dim(V)$.
- If $\dim(S) = \dim(V) < \infty$, then $S = V$.

Direct Sums

Let V be a vector space, and let $S \subseteq V$ and $T \subseteq V$ be subspaces.

We say V is the **direct sum** of S and T, written $V = S \oplus T$, if and only if for every $\boldsymbol{v} \in V$, there exist unique $\boldsymbol{s} \in S$ and $\boldsymbol{t} \in T$ such that $\boldsymbol{v} = \boldsymbol{s} + \boldsymbol{t}$.

If $V = S \oplus T$, then T is called a **complement** of S.

Example:

$$V = C' = \{f : \mathbb{R} \to \mathbb{R} | f \text{is continuous}\}$$

$$S = \text{even funcitons in} C'$$

$$T = \text{odd funcitons in} C'$$

$$f(t) = \frac{1}{2}\left(f(t) + f(-t)\right) + \frac{1}{2}\left(f(t) - f(-t)\right)$$

If $f = g + h = g' + h'$, $g \in S$ and $g' \in S$, $h \in T$ and $h' \in T$, then $g - g' = h' - h$ is odd and even, which implies $g = g'$ and $h = h'$.

Facts

1. Every subspace has a complement
2. $V = S \oplus T$ if and only if
    1. $S \cap T = \{\mathbf{0}\}$
    2. $< S, T >= V$

3. If $V = S \oplus T$, and $\dim\left(V\right) < \infty$, then $\dim\left(V\right) = \dim\left(S\right) + \dim\left(T\right)$

Proofs

Invoke a basis.

Norms

Let V be a vector space over F. A norm is a mapping $V \rightarrow F$, denoted by $\| \cdot \|$, such that forall $\boldsymbol{u} \in V$, $\boldsymbol{v} \in V$, and $\lambda \in F$

1. $\| \boldsymbol{u} \| > 0$ if $\boldsymbol{u} \neq \mathbf{0}$
2. $\| \lambda\boldsymbol{u} \| = |\lambda| \; \| \boldsymbol{u} \|$
3. $\| \boldsymbol{u} + \boldsymbol{v} \| \leq \| \boldsymbol{u} \| + \| \boldsymbol{v} \|$

Examples

Euclidean norms:

$\boldsymbol{x} \in \mathbb{R}^N$:

$$\| \boldsymbol{x} \| = \left(\sum_{i=1}^{N} {x_i}^2\right)^{\frac{1}{2}}$$

$\boldsymbol{x} \in \mathbb{C}^N$:

$$\| \boldsymbol{x} \| = \left(\sum_{i=1}^{N} \left(|x_i|\right)^2\right)^{\frac{1}{2}}$$

Induced Metric

Every norm induces a metric on V

$$d(\boldsymbol{u},\boldsymbol{v}) \equiv \| \boldsymbol{u}-\boldsymbol{v} \|$$

which leads to a notion of "distance" between vectors.

Inner products

Let V be a vector space over F, $F=\mathbb{R}$ or $\mathbb{C}$. An inner product is a mapping $V \times V \rightarrow F$, denoted $\langle \cdot, \cdot \rangle$, such that

1. $\langle \boldsymbol{v},\boldsymbol{v} \rangle \geq 0$, and $\langle \boldsymbol{v},\boldsymbol{v} \rangle = 0 \Leftrightarrow \boldsymbol{v}=0$
2. $\langle \boldsymbol{u},\boldsymbol{v} \rangle = \overline{\langle \boldsymbol{v},\boldsymbol{u} \rangle}$
3. $\langle a\boldsymbol{u}+b\boldsymbol{v},\boldsymbol{w} \rangle = a\left\langle (\boldsymbol{u},\boldsymbol{w}) \right\rangle + b\left\langle (\boldsymbol{v},\boldsymbol{w}) \right\rangle$

Examples

$\mathbb{R}^N$ over $\mathbb{R}$:

$$\langle \boldsymbol{x},\boldsymbol{y} \rangle = \left(\boldsymbol{x}^T\boldsymbol{y}\right) = \sum_{i=1}^{N} x_i y_i$$

$\mathbb{C}^N$ over $\mathbb{C}$:

$$\langle \boldsymbol{x},\boldsymbol{y} \rangle = \left(\boldsymbol{x}^{\mathrm{H}}\boldsymbol{y}\right) = \sum_{i=1}^{N} \overline{x_i} y_i$$

If $\left(\boldsymbol{x}=(x_1 ... x_N)^T\right) \in \mathbb{C}$, then

$$\boldsymbol{x}^{\mathrm{H}} \equiv \begin{pmatrix} \overline{x_1} \\ \vdots \\ \overline{x_N} \end{pmatrix}^T$$

is called the "Hermitian," or "conjugate transpose" of $\boldsymbol{x}$.

Triangle Inequality

If we define $\parallel \boldsymbol{u} \parallel = \langle \boldsymbol{u}, \boldsymbol{u} \rangle$, then

$$\parallel \boldsymbol{u} + \boldsymbol{v} \parallel \leq \parallel \boldsymbol{u} \parallel + \parallel \boldsymbol{v} \parallel$$

Hence, every inner product induces a norm.

Cauchy-Schwarz Inequality

For all $\boldsymbol{u} \in V, \boldsymbol{v} \in V$,

$$|\langle \boldsymbol{u}, \boldsymbol{v} \rangle| \leq \parallel \boldsymbol{u} \parallel \ \parallel \boldsymbol{v} \parallel$$

In inner product spaces, we have a notion of the angle between two vectors:

$$\left(\angle(\boldsymbol{u}, \boldsymbol{v}) = \operatorname{arccos}\left(\frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\parallel \boldsymbol{u} \parallel \ \parallel \boldsymbol{v} \parallel}\right)\right) \in [0, 2\pi)$$

Orthogonality

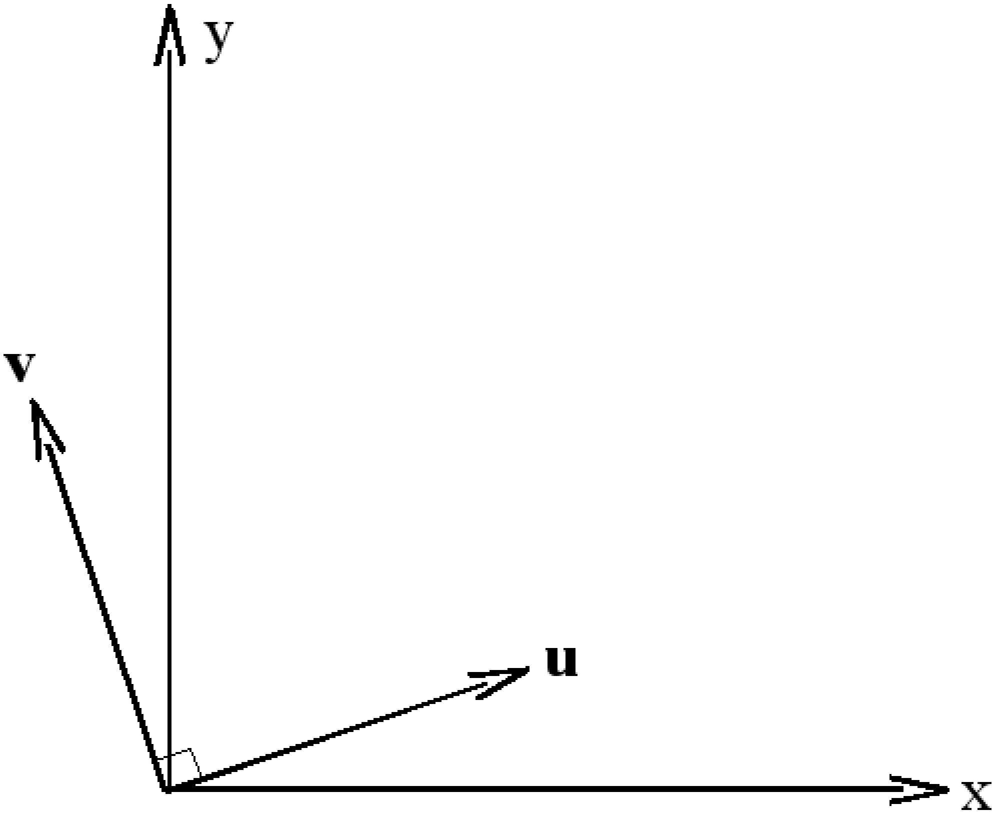
$\boldsymbol{u}$ and $\boldsymbol{v}$ are **orthogonal** if

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$$

Notation: $\boldsymbol{u} \perp \boldsymbol{v}$.

If in addition $\parallel \boldsymbol{u} \parallel = \parallel \boldsymbol{v} \parallel = 1$, we say $\boldsymbol{u}$ and $\boldsymbol{v}$ are **orthonormal**.

In an orthogonal (orthonormal) **set**, each pair of vectors is orthogonal (orthonormal).



Orthogonal vectors in $\mathbb{R}^2$.

Orthonormal Bases

An Orthonormal basis is a basis $\{v_i\}$ such that

$$\langle v_i, v_i \rangle = \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Example:
The standard basis for $\mathbb{R}^N$ or $\mathbb{C}^N$

Example:
The normalized DFT basis

$$u_k = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 \\ e^{-\left(i2\pi\frac{k}{N}\right)} \\ \vdots \\ e^{-\left(i2\pi\frac{k}{N}(N-1)\right)} \end{pmatrix}$$

Expansion Coefficients

If the representation of $\boldsymbol{v}$ with respect to $\{v_i\}$ is

$$\boldsymbol{v} = \sum_i a_i v_i$$

then

$$a_i = \langle v_i, \boldsymbol{v} \rangle$$

Gram-Schmidt

Every inner product space has an orthonormal basis. Any (countable) basis can be made orthogonal by the Gram-Schmidt orthogonalization process.

Orthogonal Compliments

Let $S \subseteq V$ be a subspace. The **orthogonal compliment** S is

$$S^{\perp} = \{\boldsymbol{u} | \, \boldsymbol{u} \in V \ \wedge \ (\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0) \ \wedge \ \forall \boldsymbol{v} : (\boldsymbol{v} \in S)\}$$

$S^{\perp}$ is easily seen to be a subspace.

If $\dim{(v)} < \infty$, then $V = S \oplus S^{\perp}$.

Note: If $\dim{(v)} = \infty$, then in order to have $V = S \oplus S^{\perp}$ we require V to be a **Hilbert Space**.

Linear Transformations

Loosely speaking, a linear transformation is a mapping from one vector space to another that **preserves** vector space operations.

More precisely, let V, W be vector spaces over the same field F. A **linear transformation** is a mapping $T : V \to W$ such that

$$T(a\boldsymbol{u} + b\boldsymbol{v}) = aT(\boldsymbol{u}) + bT(\boldsymbol{v})$$

for all $a \in F$, $b \in F$ and $\boldsymbol{u} \in V$, $\boldsymbol{v} \in V$.

In this class we will be concerned with linear transformations between (real or complex) **Euclidean spaces**, or subspaces thereof.

Image

$$\text{image}\,(T) = \{\boldsymbol{w} | \, \boldsymbol{w} \in W \ \wedge \ T(\boldsymbol{v}) = \boldsymbol{w}\text{for some}\boldsymbol{v}\}$$

Nullspace

Also known as the kernel:

$$\ker{(T)} = \{\boldsymbol{v} | \, \boldsymbol{v} \in V \ \wedge \ (T(\boldsymbol{v}) = \boldsymbol{0})\}$$

Both the image and the nullspace are easily seen to be subspaces.

Rank

$$\mathrm{rank}\,(T) = \dim\,(\mathrm{image}\;(T))$$

Nullity

$$\mathrm{null}\,(T) = \dim\,(\ker\,(T))$$

Rank plus nullity theorem

$$\mathrm{rank}\,(T) + \mathrm{null}\,(T) = \dim\,(V)$$

Matrices

Every linear transformation T has a **matrix representation**. If $T : \mathbb{E}^N \to \mathbb{E}^M$, $\mathbb{E} = \mathbb{R}$ or $\mathbb{C}$, then T is represented by an $M \times N$ matrix

$$A = \begin{pmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1} & \dots & a_{MN} \end{pmatrix}$$

where $(a_{1i} \dots a_{Mi})^T = T(e_i)$ and $e_i = (0 \dots 1 \dots 0)^T$ is the i^{th} **standard basis** vector.

Note:A linear transformation can be represented with respect to any bases of $\mathbb{E}^N$ and $\mathbb{E}^M$, leading to a different A. We will always represent a linear transformation using the standard bases.

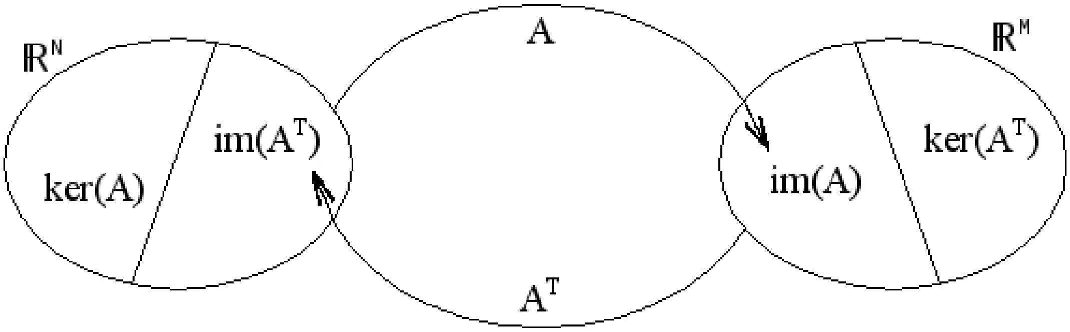
Column span

$$\text{colspan}\,(A) =< A >= \text{image}\,(A)$$

Duality

If $A : \mathbb{R}^N \to \mathbb{R}^M$, then

$$\ker^{\perp}(A) = \text{image}\left(A^T\right)$$



If $A : \mathbb{C}^N \to \mathbb{C}^M$, then

$$\ker^{\perp}(A) = \text{image}\left(A^{\mathrm{H}}\right)$$

Inverses

The linear transformation/matrix A is **invertible** if and only if there exists a matrix B such that $AB = BA = I$ (identity).

Only **square** matrices can be invertible.

Let $A : \mathbb{F}^N \to \mathbb{F}^N$ be linear, $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. The following are equivalent:

1. A is invertible (nonsingular)
2. $\text{rank}\,(A) = N$
3. $\text{null}\,(A) = 0$

4. $\det A \neq 0$
5. The columns of A form a basis.

If $A^{-1} = A^T$ (or A^H in the complex case), we say A is **orthogonal** (or **unitary**).

The Bayesian Paradigm

Statistical analysis is fundamentally an **inversion** process. The objective is to the "causes"--parameters of the probabilistic data generation model--from the "effects"--observations. This can be seen in our interpretation of the likelihood function.

Given a parameter $\boldsymbol{\theta}$, observations are generated according to

$$\mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)$$

The likelihood function has the same form as the conditional density function above

$$l(\boldsymbol{\theta}|\boldsymbol{x}) \equiv \mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)$$

except now $\boldsymbol{x}$ is given (we take measurements) and $\boldsymbol{\theta}$ is the variable. The likelihood function essentially inverts the role of observation (effect) and parameter (cause).

Unfortunately, the likelihood function does not provide a formal framework for the desired inversion.

One problem is that the parameter $\boldsymbol{\theta}$ is supposed to be a fixed and deterministic quantity while the observation $\boldsymbol{x}$ is the realization of a random process. So their role aren't really interchangeable in this setting.

Moreover, while it is tempting to interpret the likelihood $l(\boldsymbol{\theta}|\boldsymbol{x})$ as a density function for $\boldsymbol{\theta}$, this is not always possible; for example, often

$$\int l(\boldsymbol{\theta}|\boldsymbol{x})\ \mathrm{d}\ \boldsymbol{\theta} \to \infty$$

Another problematic issue is the mathematical formalization of statements like: "Based on the measurements $\boldsymbol{x}$, I am 95% confident that $\boldsymbol{\theta}$ falls in a certain range."

Example:
Suppose you toss a coin 10 times and each time it comes up "heads." It might be reasonable to say that we are 99% sure that the coin is unfair, biased towards heads.

Formally:

$$H_0 : \theta \equiv \text{prob heads} > 0.5$$

$$\boldsymbol{x} \sim \binom{N}{\sum x}\theta^{\sum x}(1-\theta)^{N-\sum x}$$

which is the binomial likelihood.

$$\mathrm{p}\left(\theta > 0.5 \,|\, \boldsymbol{x}\right) = \textit{?}$$

The problem with this is that

$$\mathrm{p}\left(\boldsymbol{\theta} \in H_0 \,|\, \boldsymbol{x}\right)$$

implies that $\boldsymbol{\theta}$ is a **random**, not deterministic, quantity. So, while "confidence" statements are very reasonable and in fact a normal part of "everyday thinking," this idea can not be supported from the classical perspective.

All of these "deficiencies" can be circumvented by a change in how we view the parameter $\boldsymbol{\theta}$.

If we view $\boldsymbol{\theta}$ as the realization of a random variable with density $\mathrm{p}\left(\boldsymbol{\theta}\right)$, then **Bayes Rule** (Bayes, 1763) shows that

$$\mathrm{p}\left(\boldsymbol{\theta} \,|\, \boldsymbol{x}\right) = \frac{\mathrm{p}\left(\boldsymbol{x} \,|\, \boldsymbol{\theta}\right) \mathrm{p}\left(\boldsymbol{\theta}\right)}{\int \mathrm{p}\left(\boldsymbol{x} \,|\, \boldsymbol{\theta}\right) \mathrm{p}\left(\boldsymbol{\theta}\right) \mathrm{d}\, \boldsymbol{\theta}}$$

Thus, from this perspective we obtain a well-defined inversion: Given $\boldsymbol{x}$, the parameter $\boldsymbol{\theta}$ is distributing according to $\mathrm{p}\left(\boldsymbol{\theta} \,|\, \boldsymbol{x}\right)$.

From here, confidence measures such as $\mathrm{p}\left(\boldsymbol{\theta} \in H_0 \,|\, \boldsymbol{x}\right)$ are perfectly legitimate quantities to ask for.

Bayesian statistical model

> A statistical model compose of a data generation model, $\mathrm{p}\left(\boldsymbol{x} \,|\, \boldsymbol{\theta}\right)$, and a **prior** distribution on the parameters, $\mathrm{p}\left(\boldsymbol{\theta}\right)$.

The **prior distriubtion** (or **prior** for short) models the uncertainty in the parameter. More specifically, $\mathrm{p}\left(\boldsymbol{\theta}\right)$ models our knowledge--or lack thereof--prior to collecting data.

Notice that

$$\mathrm{p}\left(\boldsymbol{\theta}|\boldsymbol{x}\right)=\frac{\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\mathrm{p}\left(\boldsymbol{\theta}\right)}{\mathrm{p}\left(\boldsymbol{x}\right)}\propto\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\mathrm{p}\left(\boldsymbol{\theta}\right)$$

since the data $\boldsymbol{x}$ are **known**, $\mathrm{p}\left(\boldsymbol{x}\right)$ is just a constant. Hence, $\mathrm{p}\left(\boldsymbol{\theta}|\boldsymbol{x}\right)$ is proportional to the likelihood function multiplied by the prior.

Bayesian analysis has some significant advantages over classical statistical analysis:

1. properly inverts the relationship between causes and effects
2. permits meaningful assessments in confidence regions
3. enables the incorporation of prior knowledge into the analysis (which could come from previous experiments, for example)
4. leads to more accurate estimators (provided the prior knowledge is accurate)
5. obeys the Likelihood and Sufficiency principles

Example:

$$\forall n, n=\{1,\ldots,N\}:\left(x_n=A+W_n\right)$$

$$W_n\sim\mathcal{N}\left(0,\sigma^2\right)$$

iid.

$$\widehat{A}=\frac{1}{N}\sum_{n=1}^{N}x_n$$

MVUB and MLE estimator. Now suppose that we have prior knowledge that $-A_0\leq A\leq A_0$. We might incorporate this by forming a new estimator

Equation:

$$\widetilde{A}=\begin{cases}-A_0 \text{ if } \widehat{A}<-A_0\\ \widehat{A} \text{ if } -A_0\leq\widehat{A}\leq A_0\\ A_0 \text{ if } \widehat{A}>A_0\end{cases}$$

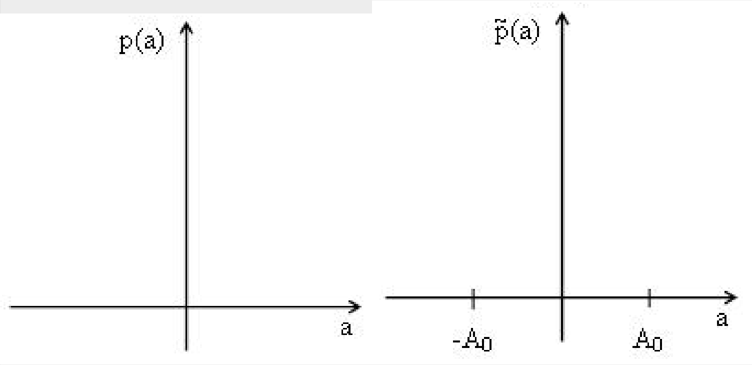
This is called a **truncated** sample mean estimator of A. Is $\widetilde{A}$ a better estimator of A than the sample mean $\widehat{A}$?

Let $\mathrm{p}\ (a)$ denote the density of $\widehat{A}$. Since $\widehat{A} = \frac{1}{N} \sum x_n$, $\mathrm{p}\ (a) = \mathcal{N}\left(A, \frac{\sigma^2}{N}\right)$.

The density of $\widetilde{A}$ is given by

Equation:

$$\widetilde{p}(a) = \Pr\left[\widehat{A} \le -A_0\right]\delta(a+A_0) + \mathrm{p}\ (a) I_{\{-A_0 \le \alpha \le A_0\}} + \Pr\left[\widehat{A} \ge A_0\right]\delta(a-A_0)$$



Now consider the MSE of the sample mean $\widehat{A}$.

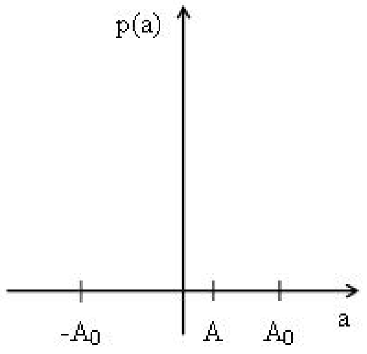
Equation:

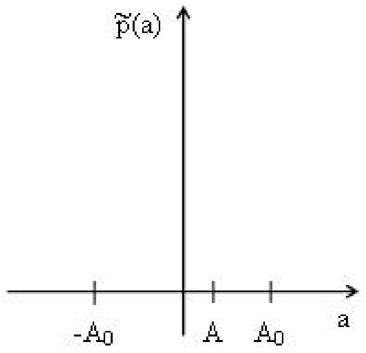
$$\mathrm{MSE}\left(\widehat{A}\right) = \int_{-\infty}^{\infty} (a-A)^2\ \mathrm{p}\ (a)\ \mathrm{d}\ a$$

Note

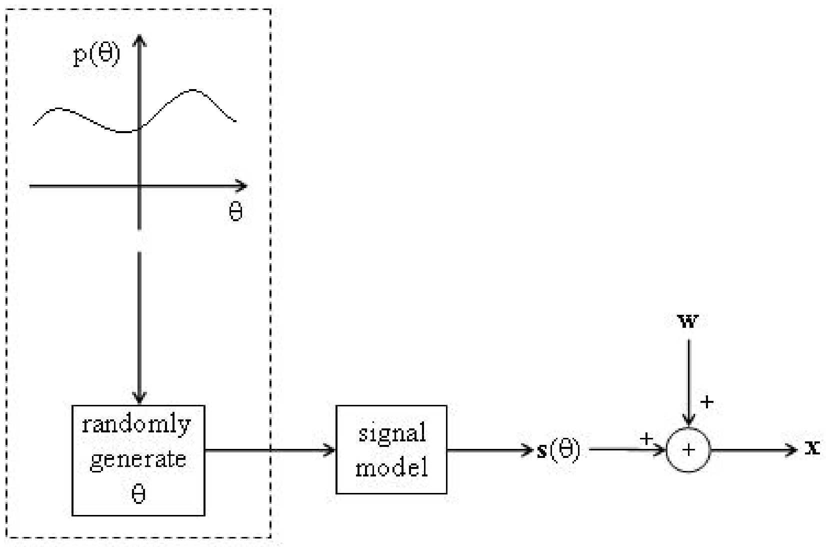
1. $\widetilde{A}$ is biased ([link]).
2. Although $\widehat{A}$ is MVUB, $\widetilde{A}$ is better in the MSE sense.
3. Prior information is aptly described by regarding A as a random variable with a prior distribution $U(-A_0, A_0)$, which implies that we know $-A_0 \le A \le A_0$, but otherwise A is abitrary.

Mean of $\widehat{A} = A$. Mean of $\widetilde{A} \neq A$.





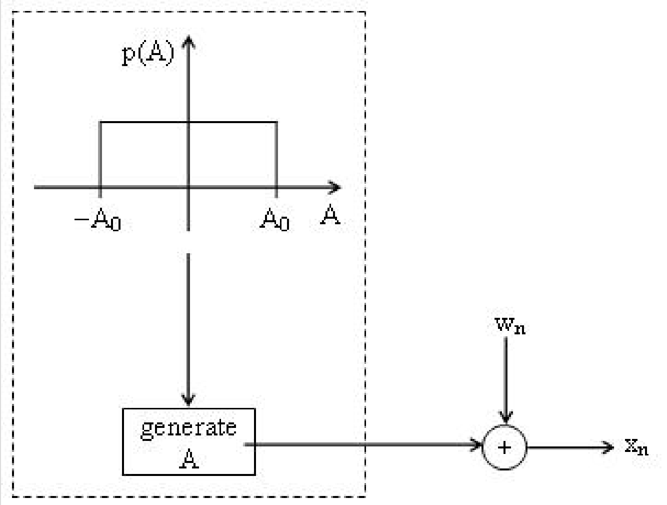
The Bayesian Approach to Statistical Modeling



Where $\boldsymbol{w}$ is the noise and $\boldsymbol{x}$ is the observation.

Example:

$$\forall n, n = \{1, \ldots, N\} : (x_n = A + W_n)$$



Prior distribution allows us to incorporate prior information regarding unknown paremter--probable values of parameter are supported by prior. Basically, the prior reflects what we believe "Nature" will probably throw at us.

Elements of Bayesian Analysis

- **(a)** joint distribution

$$\mathrm{p}\left(\boldsymbol{x},\boldsymbol{\theta}\right)=\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\mathrm{p}\left(\boldsymbol{\theta}\right)$$

- **(b)** marginal distributions

$$\mathrm{p}\left(\boldsymbol{x}\right)=\int\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\mathrm{p}\left(\boldsymbol{\theta}\right)\mathrm{d}\,\boldsymbol{\theta}$$

$$\mathrm{p}\left(\boldsymbol{\theta}\right)=\int\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\mathrm{p}\left(\boldsymbol{\theta}\right)\mathrm{d}\,\boldsymbol{x}$$

  where $\mathrm{p}\left(\boldsymbol{\theta}\right)$ is a **prior**.
- **(c)** posterior distribution

$$\mathrm{p}\left(\boldsymbol{\theta}|\boldsymbol{x}\right)=\frac{\mathrm{p}\left(\boldsymbol{x},\boldsymbol{\theta}\right)}{\mathrm{p}\left(\boldsymbol{x}\right)}=\frac{\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\mathrm{p}\left(\boldsymbol{\theta}\right)}{\int\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\mathrm{p}\left(\boldsymbol{\theta}\right)\mathrm{d}\,\boldsymbol{x}}$$

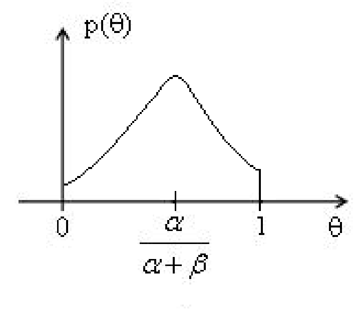
Example:

$$\forall\theta,\theta\in[0,1]:\left(\mathrm{p}\ (\boldsymbol{x}\,|\,\boldsymbol{\theta})=\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\right)$$

which is the Binomial likelihood.

$$\mathrm{p}\ (\boldsymbol{\theta})=\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

which is the Beta prior distriubtion and $B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$



This reflects prior knowledge that most probable values of θ are close to $\frac{\alpha}{\alpha+\beta}$.

Joint Density

$$\mathrm{p}\ (\boldsymbol{x},\boldsymbol{\theta})=\frac{\binom{n}{x}}{B(\alpha,\beta)}\theta^{\alpha+x-1}(1-\theta)^{n-x+\beta-1}$$

marginal density

$$\mathrm{p}\ (\boldsymbol{x})=\binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+x)\Gamma(n-\mathrm{x}+\beta)}{\Gamma(\alpha+\beta+n)}$$

posterior density

$$\mathrm{p}\left(\boldsymbol{\theta}|\boldsymbol{x}\right)=\frac{\theta^{\alpha+x-1}\theta^{\beta+n-x-1}}{B(\alpha+x,\beta+n-x)}$$

where $B(\alpha+x,\beta+n-x)$ is the Beta density with parameters $\alpha'=\alpha+x$ and $\beta'=\beta+n-x$

Selecting an Informative Prior

Clearly, the most important objective is to choose the prior $\mathrm{p}\left(\theta\right)$ that best reflects the prior knowledge available to us. In general, however, our prior knowledge is imprecise and any number of prior densities may aptly capture this information. Moreover, usually the optimal estimator can't be obtained in closed-form.

Therefore, sometimes it is desirable to choose a prior density that models prior knowledge **and** is nicely matched in functional form to $\mathrm{p}\left(\boldsymbol{x}|\theta\right)$ so that the optimal esitmator (and posterior density) can be expressed in a simple fashion.

Choosing a Prior

1. Informative Priors

- design/choose priors that are compatible with prior knowledge of unknown parameters

2. Non-informative Priors

- attempt to remove subjectiveness from Bayesian procedures
- designs are often based on invariance arguments

Example:
Suppose we want to estimate the variance of a process, incorporating a prior that is amplitude-scale invariant (so that we are invariant to arbitrary amplitude

rescaling of data). $\mathrm{p}\left(s\right) = \frac{1}{s}$ satisifies this condition. $\sigma^2 \sim \mathrm{p}\left(s\right) \Rightarrow A\sigma^2 \sim \mathrm{p}\left(s\right)$ where $\mathrm{p}\left(s\right)$ is non-informative since it is invariant to amplitude-scale.

Conjugate Priors

Idea

Given $\mathrm{p}\left(\boldsymbol{x}|\theta\right)$, choose $\mathrm{p}\left(\theta\right)$ so that $\mathrm{p}\left(\theta|\boldsymbol{x}\right) \propto \mathrm{p}\left(\boldsymbol{x}|\theta\right)\mathrm{p}\left(\theta\right)$ has a simple functional form.

Conjugate Priors

Choose $\mathrm{p}\left(\theta\right) \in \mathscr{P}$, where $\mathscr{P}$ is a family of densities (e.g., Gaussian family) so that the posterior density also belongs to that family.

conjugate prior

$\mathrm{p}\left(\theta\right)$ is a **conjugate prior** for $\mathrm{p}\left(\boldsymbol{x}|\theta\right)$ if $\mathrm{p}\left(\theta\right) \in \mathscr{P} \Rightarrow \mathrm{p}\left(\theta|\boldsymbol{x}\right) \in \mathscr{P}$

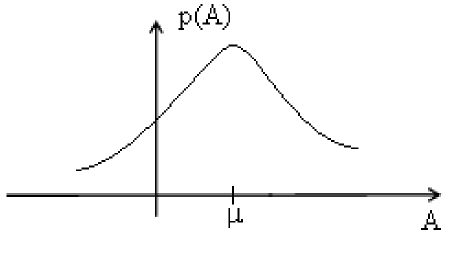
Example:

$$\forall n, n = \{1, \ldots, N\} : (x_n = A + W_n)$$

$$W_n \sim \mathscr{N}\left(0, \sigma^2\right)$$

iid. Rather than modeling $A \sim U(-A_0, A_0)$ (which did not yield a closed-form estimator) consider

$$\mathrm{p}\left(A\right) = \frac{1}{\sqrt{2\pi{\sigma_A}^2}} e^{\frac{-1}{2{\sigma_A}^2}(A-\mu)^2}$$



With $\mu = 0$ and $\sigma_A = \frac{1}{3} A_0$ this Gaussian prior also reflects prior knowledge that it is unlikely for $|A| \geq A_0$.

The Gaussian prior is also conjugate to the Gaussian likelihood

$$\mathrm{p}\left(\boldsymbol{x}\,|\,A\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}\left(x_n - A\right)^2}$$

so that the resulting posterior density is also a simple Gaussian, as shown next.

First note that

$$\mathrm{p}\left(\boldsymbol{x}\,|\,A\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N} x_n} e^{\frac{-1}{2\sigma^2}\left(NA^2 - 2NA\bar{x}\right)}$$

where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$.

Equation:

$$\begin{aligned}
\mathrm{p}\left(A\,|\,\boldsymbol{x}\right) &= \frac{\mathrm{p}(\boldsymbol{x}\,|\,A)\mathrm{p}(A)}{\int \mathrm{p}(\boldsymbol{x}\,|\,A)\mathrm{p}(A)\mathrm{d}A} \\
&= \frac{e^{\frac{-1}{2}\left(\frac{1}{\sigma^2}\left(NA^2 - 2NA\bar{x}\right) + \frac{1}{{\sigma_A}^2}(A-\mu)^2\right)}}{\int_{-\infty}^{\infty} e^{\frac{-1}{2}\left(\frac{1}{\sigma^2}\left(NA^2 - 2NA\bar{x}\right) + \frac{1}{{\sigma_A}^2}(A-\mu)^2\right)}\mathrm{d}A} \\
&= \frac{e^{\frac{-1}{2}Q(A)}}{\int_{-\infty}^{\infty} e^{\frac{-1}{2}Q(A)}\mathrm{d}A}
\end{aligned}$$

where $Q(A) = \frac{N}{\sigma^2}A^2 - \frac{2NA\bar{x}}{\sigma^2} + \frac{A^2}{{\sigma_A}^2} - \frac{2\mu A}{{\sigma_A}^2} + \frac{\mu^2}{{\sigma_A}^2}$. Now let

$$\sigma_{A|x}{}^{2} \equiv \frac{1}{\frac{N}{\sigma^{2}}+\frac{1}{\sigma_{A}{}^{2}}}$$

$$\mu_{A|x}{}^{2} \equiv \left(\frac{N}{\sigma^{2}}\bar{x}+\frac{\mu}{\sigma_{A}{}^{2}}\right)\sigma_{A|x}{}^{2}$$

Then by "completing the square" we have
Equation:

$$\begin{aligned} Q(A) &= \tfrac{1}{\sigma_{A|x}{}^{2}}\left(A^{2}-2\mu_{A|x}A+\mu_{A|x}{}^{2}\right)-\tfrac{\mu_{A|x}{}^{2}}{\sigma_{A|x}{}^{2}}+\tfrac{\mu^{2}}{\sigma_{A}{}^{2}} \\ &= \tfrac{1}{\sigma_{A|x}{}^{2}}\left(A-\mu_{A|x}\right)^{2}-\tfrac{\mu_{A|x}{}^{2}}{\sigma_{A|x}{}^{2}}+\tfrac{\mu^{2}}{\sigma_{A}{}^{2}} \end{aligned}$$

Hence,

Math input error

where Math input error is the "unnormalized" Gaussian density and $\frac{-1}{2}\left(\frac{\mu^{2}}{\sigma_{A}{}^{2}}-\frac{\mu_{A|x}{}^{2}}{\sigma_{A|x}{}^{2}}\right)$ is a constant, independent of A. This implies that

$$\mathrm{p}\left(A\,|\,\boldsymbol{x}\right)=\frac{1}{\sqrt{2\pi\sigma_{A|x}{}^{2}}}e^{\frac{-1}{2\sigma_{A|x}{}^{2}}\left(A-\mu_{A|x}\right)^{2}}$$

where $A|\boldsymbol{x} \sim \mathcal{N}\left(\mu_{A|x},\sigma_{A|x}{}^{2}\right)$. Now
Equation:

$$\begin{aligned} \widehat{A} &= E[A\,|\,\boldsymbol{x}] \\ &= \int A\,\mathrm{p}\left(A\,|\,\boldsymbol{x}\right)\mathrm{d}\,A \\ &= \mu_{A|x} \\ &= \frac{\frac{N}{\sigma^{2}}\bar{x}+\frac{\mu}{\sigma_{A}{}^{2}}}{\frac{N}{\sigma^{2}}+\frac{1}{\sigma_{A}{}^{2}}} \\ &= \frac{\sigma_{A}{}^{2}}{\sigma_{A}{}^{2}+\frac{\sigma^{2}}{N}}\bar{\boldsymbol{x}}+\frac{\frac{\sigma^{2}}{N}}{\sigma_{A}{}^{2}+\frac{\sigma^{2}}{N}}\mu \\ &= \alpha\bar{\boldsymbol{x}}-1\mu \end{aligned}$$

Where $0 < \alpha = \frac{{\sigma_A}^2}{{\sigma_A}^2+\frac{\sigma^2}{N}} < 1$

Interpretation

1. When there is little data ${\sigma_A}^2 \ll \frac{\sigma^2}{N}$ α is small and $\widehat{A} = \mu$.
2. When there is a lot of data ${\sigma_A}^2 \gg \frac{\sigma^2}{N}$, $\alpha \simeq 1$ and $\widehat{A} = \bar{x}$.

Interplay Between Data and Prior Knowledge

Small $N \rightarrow \widehat{A}$ favors prior.

Large $N \rightarrow \widehat{A}$ favors data.

The Multivariate Gaussian Model

The multivariate Gaussian model is the most important Bayesian tool in signal processing. It leads directly to the celebrated Wiener and Kalman filters.

Assume that we are dealing with random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. We will regard $\boldsymbol{y}$ as a signal vector that is to be estimated from an observation vector $\boldsymbol{x}$.

$\boldsymbol{y}$ plays the same role as $\boldsymbol{\theta}$ did in earlier discussions. We will assume that $\boldsymbol{y}$ is p×1 and $\boldsymbol{x}$ is N×1. Furthermore, assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are **jointly** Gaussian distributed

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} R_{\text{xx}} & R_{\text{xy}} \\ R_{\text{yx}} & R_{\text{yy}} \end{pmatrix} \right)$$

$E[\boldsymbol{x}] = \mathbf{0}$, $E[\boldsymbol{y}] = \mathbf{0}$, $E\left[\boldsymbol{x}\boldsymbol{x}^T\right] = R_{\text{xx}}$, $E\left[\boldsymbol{x}\boldsymbol{y}^T\right] = R_{\text{xy}}$, $E\left[\boldsymbol{y}\boldsymbol{x}^T\right] = R_{\text{yx}}$, $E\left[\boldsymbol{y}\boldsymbol{y}^T\right] = R_{\text{yy}}$.

$$R \equiv \begin{pmatrix} R_{\text{xx}} & R_{\text{xy}} \\ R_{\text{yx}} & R_{\text{yy}} \end{pmatrix}$$

Example:
$\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{W}$, $\boldsymbol{W} \sim \mathcal{N}\left(0, \sigma^2 I\right)$

$$\text{p}\ (\boldsymbol{y}) = \mathcal{N}(\mathbf{0}, R_{\text{yy}})$$

which is independent of $\boldsymbol{W}$. $E[\boldsymbol{x}] = E[\boldsymbol{y}] + E[\boldsymbol{W}] = 0$,
$E\left[\boldsymbol{x}\boldsymbol{x}^T\right] = E\left[\boldsymbol{y}\boldsymbol{y}^T\right] + E\left[\boldsymbol{y}\boldsymbol{W}^T\right] + E\left[\boldsymbol{W}\boldsymbol{y}^T\right] + E\left[\boldsymbol{W}\boldsymbol{W}^T\right] = R_{\text{yy}} + \sigma^2 I$,
$E\left[\boldsymbol{x}\boldsymbol{y}^T\right] = E\left[\boldsymbol{y}\boldsymbol{y}^T\right] + E\left[\boldsymbol{W}\boldsymbol{y}^T\right] = R_{\text{yy}} = E\left[\boldsymbol{y}\boldsymbol{x}^T\right]$.

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \sim \mathscr{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} R_{\text{yy}} + \sigma^2 I & R_{\text{yy}} \\ R_{\text{yy}} & R_{\text{yy}} \end{pmatrix}\right)$$

From our Bayesian perpsective, we are interested in $\mathrm{p}\left(\boldsymbol{y}|\boldsymbol{x}\right)$.
Equation:

$$\begin{aligned} \mathrm{p}\left(\boldsymbol{y}|\boldsymbol{x}\right) &= \frac{\mathrm{p}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{p}(\boldsymbol{x})} \\ &= \frac{(2\pi)^{-\frac{N}{2}}(2\pi)^{-\frac{p}{2}}(\det R)^{\frac{-1}{2}} e^{\frac{-1}{2}\left(\boldsymbol{x}^T \quad \boldsymbol{y}^T\right) R^{-1} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}}}{(2\pi)^{-\frac{N}{2}}(\det R_{\text{xx}})^{\frac{-1}{2}} e^{\frac{-1}{2}\boldsymbol{x}^T {R_{\text{xx}}}^{-1}\boldsymbol{x}}} \end{aligned}$$

In this formula we are faced with

$$R^{-1} = \begin{pmatrix} R_{\text{xx}} & R_{\text{xy}} \\ R_{\text{yx}} & R_{\text{yy}} \end{pmatrix}^{-1}$$

The inverse of this covariance matrix can be written as

$$\begin{pmatrix} R_{\text{xx}} & R_{\text{xy}} \\ R_{\text{yx}} & R_{\text{yy}} \end{pmatrix}^{-1} = \begin{pmatrix} {R_{\text{xx}}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \left(-{R_{\text{xx}}}^{-1}\right) R_{\text{xy}} \\ I \end{pmatrix} Q^{-1}\left(\left(-R_{\text{yx}}\right) R_{\text{xx}} \quad I\right)$$

where $Q \equiv R_{\text{yy}} - R_{\text{yx}} R_{\text{xx}} R_{\text{xy}}$. (Verify this formula by applying the right hand side above to R to get I.)

Sufficient Statistics

Introduction

Sufficient statistics arise in nearly every aspect of statistical inference. It is important to understand them before progressing to areas such as hypothesis testing and parameter estimation.

Suppose we observe an N-dimensional random vector $\boldsymbol{X}$, characterized by the density or mass function $\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)$, where $\boldsymbol{\theta}$ is a p-dimensional vector of parameters to be estimated. The functional form of $\mathrm{f}\left(x\right)$ is assumed known. The parameter $\boldsymbol{\theta}$ completely determines the distribution of $\boldsymbol{X}$. Conversely, a measurement $\boldsymbol{x}$ of $\boldsymbol{X}$ provides information about $\boldsymbol{\theta}$ through the probability law $\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)$.

Example:
Suppose $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, where $X_i \sim \mathcal{N}(\theta, 1)$ are IID. Here θ is a scalar parameter specifying the mean. The distribution of $\boldsymbol{X}$ is determined by θ through the density

$$\mathrm{f}_{\theta}\left(\boldsymbol{x}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1-\theta)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_2-\theta)^2}{2}}$$

On the other hand, if we observe $\boldsymbol{x} = \begin{pmatrix} 100 \\ 102 \end{pmatrix}$, then we may safely assume $\theta = 0$ is highly unlikely.

The N-dimensional observation $\boldsymbol{X}$ carries information about the p-dimensional parameter vector $\boldsymbol{\theta}$. If $p < N$, one may ask the following question: Can we compress $\boldsymbol{x}$ into a low-dimensional statistic without any loss of information? Does there exist some function $\boldsymbol{t} = T(\boldsymbol{x})$, where the dimension of $\boldsymbol{t}$ is $M < N$, such that $\boldsymbol{t}$ carries all the useful information about $\boldsymbol{\theta}$?

If so, for the purpose of studying $\boldsymbol{\theta}$ we could discard the raw measurements $\boldsymbol{x}$ and retain only the low-dimensional statistic $\boldsymbol{t}$. We call $\boldsymbol{t}$ a **sufficient statistic**. The following definition captures this notion precisely:

> Let $X_1, \ldots, X_M$ be a random sample, governed by the density or probability mass function $\mathrm{f}\ (\boldsymbol{x}\,|\,\boldsymbol{\theta})$. The statistic $T(\boldsymbol{x})$ is **sufficient** for $\boldsymbol{\theta}$ if the conditional distribution of $\boldsymbol{x}$, given $T(\boldsymbol{x}) = \boldsymbol{t}$, is independent of $\boldsymbol{\theta}$. Equivalently, the functional form of $\mathrm{f}_{\boldsymbol{\theta}|\boldsymbol{t}}\ (\boldsymbol{x})$ does not involve $\boldsymbol{\theta}$.

How should we interpret this definition? Here are some possibilities:

1. Let $\mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{x}, \boldsymbol{t})$ denote the joint density or probability mass function on $(\boldsymbol{X}, \boldsymbol{T}(\boldsymbol{X}))$. If $T(\boldsymbol{X})$ is a sufficient statistic for $\boldsymbol{\theta}$, then
Equation:

$$\begin{aligned} \mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{x}) &= \mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{x}, T(\boldsymbol{x})) \\ &= \mathrm{f}_{\boldsymbol{\theta}|\boldsymbol{t}}\ (\boldsymbol{x})\ \mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{t}) \\ &= \mathrm{f}\ (\boldsymbol{x}\,|\,\boldsymbol{t})\ \mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{t}) \end{aligned}$$

Therefore, the parametrization of the probability law for the measurement $\boldsymbol{x}$ is manifested in the parametrization of the probability law for the statistic $T(\boldsymbol{x})$.

2. Given $\boldsymbol{t} = T(\boldsymbol{x})$, full knowledge of the measurement $\boldsymbol{x}$ brings no additional information about $\boldsymbol{\theta}$. Thus, we may discard $\boldsymbol{x}$ and retain on the compressed statistic $\boldsymbol{t}$.

3. Any inference strategy based on $\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)$ may be replaced by a strategy based on $\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{t}\right)$.

Example:

Binary Information Source

(Scharf, pp.78) Suppose a binary information source emits a sequence of binary (0 or 1) valued, independent variables $x_1, \ldots, x_N$. Each binary symbol may be viewed as a realization of a Bernoulli trial: $x_n \sim$ Bernoulli (θ), iid. The parameter $\theta \in [0, 1]$ is to be estimated.

The probability mass function for the random sample $\boldsymbol{x} = (x_1 \ldots x_N)^T$ is

Equation:

$$\mathrm{f}_{\theta}\left(\boldsymbol{x}\right) = \prod_{n=1}^{N} \mathrm{f}_{\theta}\left(x_n\right) \prod_{n=1}^{N} \theta^k (1-\theta)^{N-k}$$

where $k = \sum_{n=1}^{N} x_n$ is the number of 1's in the sample.

We will show that k is a sufficient statistic for $\boldsymbol{x}$. This will entail showing that the conditional probability mass function $\mathrm{f}_{\theta|k}\left(\boldsymbol{x}\right)$ does not depend on θ.

The distribution of the number of ones in N independent Bernoulli trials is binomial:

$$\mathrm{f}_{\theta}\left(k\right) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

Next, consider the joint distribution of $(\boldsymbol{x}, \sum x_n)$. We have

$$\mathrm{f}_{\theta}\;(\boldsymbol{x})\;=\mathrm{f}_{\theta}\left(\boldsymbol{x},\sum x_n\right)$$

Thus, the conditional probability may be written
Equation:

$$\begin{aligned}\mathrm{f}_{\theta|k}\;(\boldsymbol{x}) &= \frac{\mathrm{f}_{\theta}(\boldsymbol{x},k)}{\mathrm{f}_{\theta}(k)} \\ &= \frac{\mathrm{f}_{\theta}(\boldsymbol{x})}{\mathrm{f}_{\theta}(k)} \\ &= \frac{\theta^{k}(1-\theta)^{N-k}}{\binom{N}{k}\theta^{k}(1-\theta)^{N-k}} \\ &= \frac{1}{\binom{N}{k}}\end{aligned}$$

This shows that k is indeed a sufficient statistic for θ. The N values $x_1, \ldots, x_N$ can be replaced by the quantity k without losing information about θ.

Exercise:

Problem:

In the previous example, suppose we wish to store in memory the information we possess about θ. Compare the savings, in terms of bits, we gain by storing the sufficient statistic k instead of the full sample $x_1, \ldots, x_N$.

Determining Sufficient Statistics

In the example above, we had to guess the sufficient statistic, and work out the conditional probability by hand. In general, this will be a tedious way to go about finding sufficient statistics. Fortunately, spotting sufficient statistics can be made easier by the Fisher-Neyman Factorization Theorem.

Uses of Sufficient Statistics

Sufficient statistics have many uses in statistical inference problems. In hypothesis testing, the Likelihood Ratio Test can often be reduced to a sufficient statistic of the data. In parameter estimation, the Minimum Variance Unbiased Estimator of a parameter $\boldsymbol{\theta}$ can be characterized by sufficient statistics and the Rao-Blackwell Theorem.

Minimality and Completeness

Minimal sufficient statistics are, roughly speaking, sufficient statistics that cannot be compressed any more without losing information about the unknown parameter. **Completeness** is a technical characterization of sufficient statistics that allows one to prove minimality. These topics are covered in detail in this module.

Further examples of sufficient statistics may be found in the module on the Fisher-Neyman Factorization Theorem.

The Fisher-Neyman Factorization Theorem

Determining a sufficient statistic directly from the definition can be a tedious process. The following result can simplify this process by allowing one to spot a sufficient statistic directly from the functional form of the density or mass function.

Fisher-Neyman Factorization Theorem

Let $\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)$ be the density or mass function for the random vector $\boldsymbol{x}$, parametrized by the vector $\boldsymbol{\theta}$. The statistic $\boldsymbol{t} = T(\boldsymbol{x})$ is sufficient for $\boldsymbol{\theta}$ if and only if there exist functions $a(\boldsymbol{x})$ (not depending on $\boldsymbol{\theta}$) and $\mathrm{b}_{\boldsymbol{\theta}}\left(\boldsymbol{t}\right)$ such that

$$\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = a(\boldsymbol{x})\ \mathrm{b}_{\boldsymbol{\theta}}\left(\boldsymbol{t}\right)$$

for all possible values of $\boldsymbol{x}$.

In an earlier example we computed a sufficient statistic for a binary communication source (independent Bernoulli trials) from the definition. Using the above result, this task becomes substantially easier.

Example:

Bernoulli Trials Revisited

Suppose $x_n \sim \mathrm{Bernoulli}\left(\theta\right)$ are IID, $\forall n, n = 1, \ldots, N : (n = 1, \ldots, N)$. Denote $\boldsymbol{x} = \left(x_1 \ldots x_n\right)^T$. Then

Equation:

$$\begin{aligned} \mathrm{f}_{\theta}\left(\boldsymbol{x}\right) &= \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} \\ &= \theta^{k}(1-\theta)^{N-k} \\ &= a(\boldsymbol{x})\ \mathrm{b}_{\theta}\left(k\right) \end{aligned}$$

where $k=\sum_{n=1}^{N} x_n$, $a(\boldsymbol{x})=1$, and $\mathrm{b}_\theta\left(k\right)=\theta^k(1-\theta)^{N-k}$. By the Fisher-Neyman factorization theorem, k is sufficient for θ.

The next example illustrates the appliction of the theorem to a continuous random variable.

Example:

Normal Data with Unknown Mean

Consider a normally distributed random sample $x_1, \ldots, x_N \sim \mathcal{N}(\theta, 1)$, IID, where θ is unknown. The joint pdf of $\boldsymbol{x}=(x_1 \ldots x_n)^T$ is

$$\mathrm{f}_\theta\left(\boldsymbol{x}\right)=\prod_{n=1}^{N} \mathrm{f}_\theta\left(x_n\right)=\left(\frac{1}{2\pi}\right)^{\frac{N}{2}} e^{\frac{-1}{2} \sum_{n=1}^{N} (x_n-\theta)^2}$$

We would like to rewrite $\mathrm{f}_\theta\left(\boldsymbol{x}\right)$ is the form of $a(\boldsymbol{x})\ \mathrm{b}_\theta\left(\boldsymbol{t}\right)$, where dim $\left(\boldsymbol{t}\right)<N$. At this point we require a trick-one that is commonly used when manipulating normal densities, and worth remembering. Define $\bar{x}=\frac{1}{N}\sum_{n=1}^{N} x_n$, the sample mean. Then

Equation:

$$\begin{aligned}\mathrm{f}_\theta\left(\boldsymbol{x}\right) &= \left(\frac{1}{2\pi}\right)^{\frac{N}{2}} e^{\frac{-1}{2}\sum_{n=1}^{N}\left(x_n-\bar{x}+\bar{x}-\theta\right)^2} \\ &= \left(\frac{1}{2\pi}\right)^{\frac{N}{2}} e^{\frac{-1}{2}\sum_{n=1}^{N}\left(x_n-\bar{x}\right)^2+2\left(x_n-\bar{x}\right)\left(\bar{x}-\theta\right)+\left(\bar{x}-\theta\right)^2}\end{aligned}$$

Now observe
Equation:

$$\begin{aligned}\sum_{n=1}^{N}\left(x_n-\bar{x}\right)\left(\bar{x}-\theta\right) &= \left(\bar{x}-\theta\right)\sum_{n=1}^{N}x_n-\bar{x} \\ &= \left(\bar{x}-\theta\right)\left(\bar{x}-\bar{x}\right) \\ &= 0\end{aligned}$$

so the middle term vanishes. We are left with

$$\mathrm{f}_{\theta}\left(\boldsymbol{x}\right)=\left(\frac{1}{2\pi}\right)^{\frac{N}{2}}e^{\frac{-1}{2}\sum_{n=1}^{N}\left(x_n-\bar{x}\right)^2}e^{\frac{-1}{2}\sum_{n=1}^{N}\left(\bar{x}-\theta\right)^2}$$

where $a(\boldsymbol{x})=\left(\frac{1}{2\pi}\right)^{\frac{N}{2}}e^{\frac{-1}{2}\sum_{n=1}^{N}\left(x_n-\bar{x}\right)^2}$,
$\mathrm{b}_{\theta}\left(\boldsymbol{t}\right)=e^{\frac{-1}{2}\sum_{n=1}^{N}\left(\bar{x}-\theta\right)^2}$, and $\boldsymbol{t}=\boldsymbol{x}$. Thus, the sample mean is a one-dimensional sufficient statistic for the mean.

Proof of Theorem

First, suppose $\boldsymbol{t}=T(\boldsymbol{x})$ is sufficient for $\boldsymbol{\theta}$. By definition, $\mathrm{f}_{\boldsymbol{\theta}|T(\boldsymbol{x})=\boldsymbol{t}}\left(\boldsymbol{x}\right)$ is independent of $\boldsymbol{\theta}$. Let $\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x},\boldsymbol{t}\right)$ denote the joint density or mass function for $(\boldsymbol{X},\boldsymbol{T}(\boldsymbol{X}))$. Observe $\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)=\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x},\boldsymbol{t}\right)$. Then
Equation:

$$\begin{aligned}\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) &= \mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x},\boldsymbol{t}\right) \\ &= \mathrm{f}_{\boldsymbol{\theta}|\boldsymbol{t}}\left(\boldsymbol{x}\right)\mathrm{f}_{\boldsymbol{\theta}}\left(\boldsymbol{t}\right) \\ &= a(\boldsymbol{x})\,\mathrm{b}_{\boldsymbol{\theta}}\left(\boldsymbol{t}\right)\end{aligned}$$

where $a(\boldsymbol{x}) = \mathrm{f}_{\boldsymbol{\theta}|\boldsymbol{t}}\ (\boldsymbol{x})$ and $\mathrm{b}_{\boldsymbol{\theta}}\ (\boldsymbol{t}) = \mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{t})$. We prove the reverse implication for the discrete case only. The continuous case follows a similar argument, but requires a bit more technical work (Scharf, pp.82; Kay, pp.127).

Suppose the probability mass function for $\boldsymbol{x}$ can be written

$$\mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{x}) = a(\boldsymbol{x})\ \mathrm{b}_{\boldsymbol{\theta}}\ (\boldsymbol{x})$$

where $\boldsymbol{t} = T(\boldsymbol{x})$. The probability mass function for $\boldsymbol{t}$ is obtained by summing $\mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{x}, \boldsymbol{t})$ over all $\boldsymbol{x}$ such that $T(\boldsymbol{x}) = \boldsymbol{t}$:

Equation:

$$\begin{aligned} \mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{t}) &= \sum_{T(\boldsymbol{x})=\boldsymbol{t}} \mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{x}, \boldsymbol{t}) \\ &= \sum_{T(\boldsymbol{x})=\boldsymbol{t}} \mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{x}) \\ &= \sum_{T(\boldsymbol{x})=\boldsymbol{t}} a(\boldsymbol{x})\ \mathrm{b}_{\boldsymbol{\theta}}\ (\boldsymbol{t}) \end{aligned}$$

Therefore, the conditional mass function of $\boldsymbol{x}$, given $\boldsymbol{t}$, is

Equation:

$$\begin{aligned} \mathrm{f}_{\boldsymbol{\theta}|\boldsymbol{t}}\ (\boldsymbol{x}) &= \frac{\mathrm{f}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{t})}{\mathrm{f}_{\boldsymbol{\theta}}(\boldsymbol{t})} \\ &= \frac{\mathrm{f}_{\boldsymbol{\theta}}(\boldsymbol{x})}{\mathrm{f}_{\boldsymbol{\theta}}(\boldsymbol{t})} \\ &= \frac{a(\boldsymbol{x})}{\sum_{T(\boldsymbol{x})=\boldsymbol{t}} a(\boldsymbol{x})} \end{aligned}$$

This last expression does not depend on $\boldsymbol{\theta}$, so $\boldsymbol{t}$ is a sufficient statistic for $\boldsymbol{\theta}$. This completes the proof.

Note: From the proof, the Fisher-Neyman factorization gives us a formula for the conditional probability of $\boldsymbol{x}$ given $\boldsymbol{t}$. In the discrete case we have

$$\mathrm{f}\ (\boldsymbol{x}\,|\,\boldsymbol{t}) = \frac{a(\boldsymbol{x})}{\sum_{T(\boldsymbol{x})=\boldsymbol{t}} a(\boldsymbol{x})}$$

An analogous formula holds for continuous random variables (Scharf, pp.82).

Further Examples

The following exercises provide additional examples where the Fisher-Neyman factorization may be used to identify sufficient statistics.

Exercise:

Problem:
Uniform Measurements

Suppose $x_1, \ldots, x_N$ are independent and uniformly distributed on the interval $[\theta_1, \theta_2]$. Find a sufficient statistic for $\boldsymbol{\theta} = (\theta_1 \theta_2)^T$.

Note: Express the likelihood $\mathrm{f}_{\boldsymbol{\theta}}\ (\boldsymbol{x})$ in terms of indicator functions.

Exercise:

Problem:
Poisson

Suppose $x_1, \ldots, x_N$ are independent measurements of a Poisson random variable with intensity parameter θ:

$$\forall x, x = 0, 1, 2, \ldots : \left(\mathrm{f}_{\boldsymbol{\theta}}\ (x) = \frac{e^{-\theta}\theta^x}{x!}\right)$$

Find a sufficient statistic $\boldsymbol{t}$ for θ.

What is the conditional probability mass function of $\boldsymbol{x}$, given $\boldsymbol{t}$, where $\boldsymbol{x} = (x_1 \ldots x_N)^T$?

Exercise:

Problem:
Normal with Unknown Mean and Variance

Consider $x_1, \ldots, x_N \sim \mathcal{N}(\mu, \sigma^2)$, IID, where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ are both unknown. Find a sufficient statistic for $\boldsymbol{\theta} = (\theta_1 \theta_2)^T$.

Note: Use the same trick as in [link].

Signal Classifications and Properties
Describes various classifications of signals.

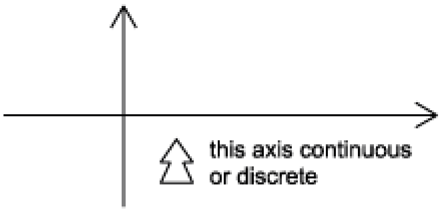
Introduction

This module will begin our study of signals and systems by laying out some of the fundamentals of signal classification. It is essentially an introduction to the important definitions and properties that are fundamental to the discussion of signals and systems, with a brief discussion of each.

Classifications of Signals

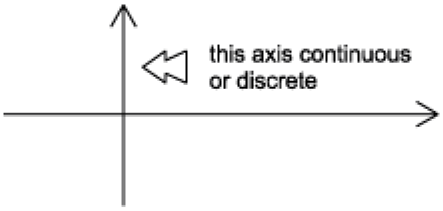
Continuous-Time vs. Discrete-Time

As the names suggest, this classification is determined by whether or not the time axis is **discrete** (countable) or **continuous** ([link]). A continuous-time signal will contain a value for all real numbers along the time axis. In contrast to this, a discrete-time signal, often created by sampling a continuous signal, will only have values at equally spaced intervals along the time axis.



Analog vs. Digital

The difference between **analog** and **digital** is similar to the difference between continuous-time and discrete-time. However, in this case the difference involves the values of the function. Analog corresponds to a continuous set of possible function values, while digital corresponds to a discrete set of possible function values. An common example of a digital signal is a binary sequence, where the values of the function can only be one or zero.



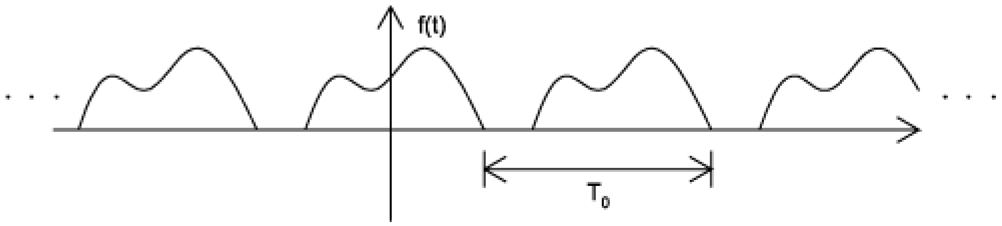
Periodic vs. Aperiodic

Periodic signals repeat with some **period** T, while aperiodic, or nonperiodic, signals do not ([link]). We can define a periodic function through the following mathematical expression, where t can be any number and T is a positive constant:

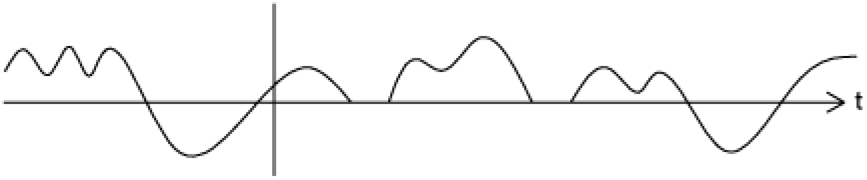
Equation:

$$f(t)=f(t+T)$$

fundamental period of our function, $f(t)$, is the smallest value of T that the still allows [link] to be true.



A periodic signal with period T_0



An aperiodic signal

Finite vs. Infinite Length

Another way of classifying a signal is in terms of its length along its time axis. Is the signal defined for all possible values of time, or for only certain values of time? Mathematically speaking, $f(t)$ is a **finite-length signal** if it is **defined** only over a finite interval

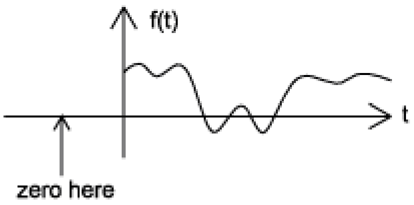
$$t_1 < t < t_2$$

where $t_1 < t_2$. Similarly, an **infinite-length signal**, $f(t)$, is defined for all values:

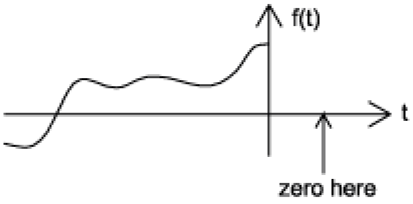
$$-\infty < t < \infty$$

Causal vs. Anticausal vs. Noncausal

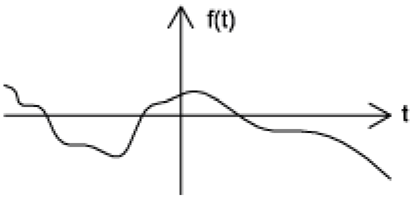
Causal signals are signals that are zero for all negative time, while **anticausal** are signals that are zero for all positive time. **Noncausal** signals are signals that have nonzero values in both positive and negative time ([link]).



A causal signal



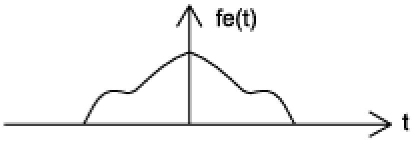
An anticausal signal



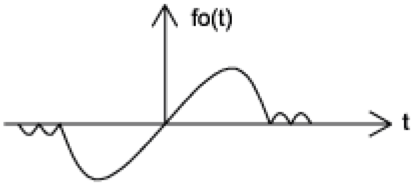
A noncausal signal

Even vs. Odd

An **even signal** is any signal f such that $f(t) = f(-t)$. Even signals can be easily spotted as they are **symmetric** around the vertical axis. An **odd signal**, on the other hand, is a signal f such that $f(t) = -f(-t)$ ([link]).



An even signal



An odd signal

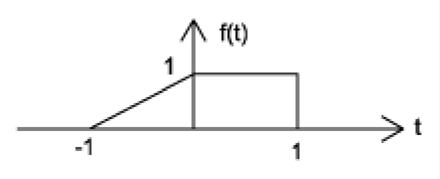
Using the definitions of even and odd signals, we can show that any signal can be written as a combination of an even and odd signal. That is, every signal has an odd-even decomposition. To demonstrate this, we have to look no further than a single equation.

Equation:

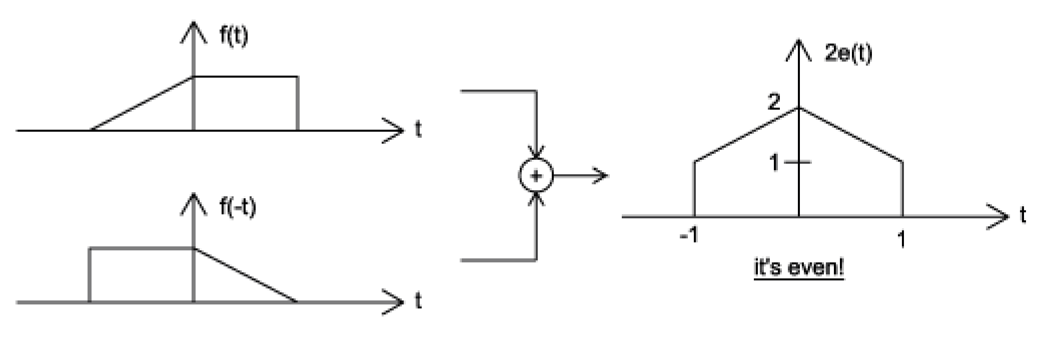
$$f(t)=\frac{1}{2}\left(f(t)+f(-t)\right)+\frac{1}{2}\left(f(t)-f(-t)\right)$$

By multiplying and adding this expression out, it can be shown to be true. Also, it can be shown that $f(t)+f(-t)$ fulfills the requirement of an even function, while $f(t)-f(-t)$ fulfills the requirement of an odd function ([link]).

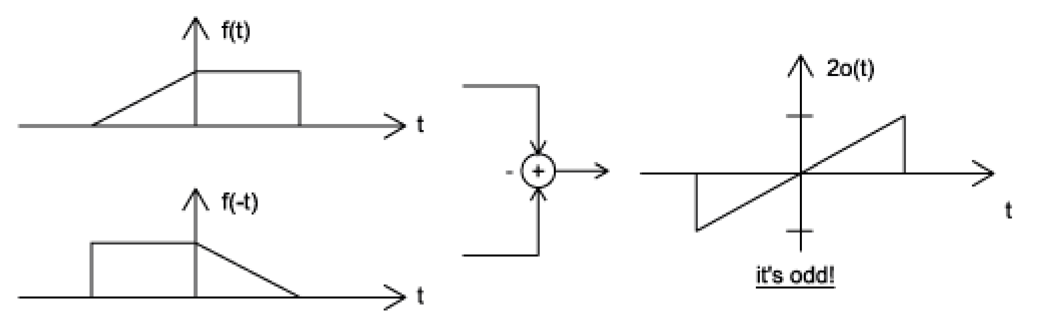
Example:



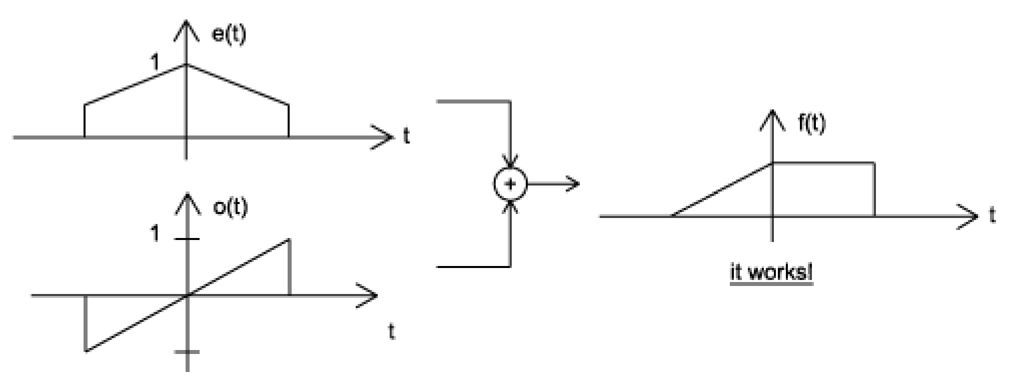
The signal we will decompose using odd-even decomposition



Even part: $e(t) = \frac{1}{2}\left(f(t) + f(-t)\right)$



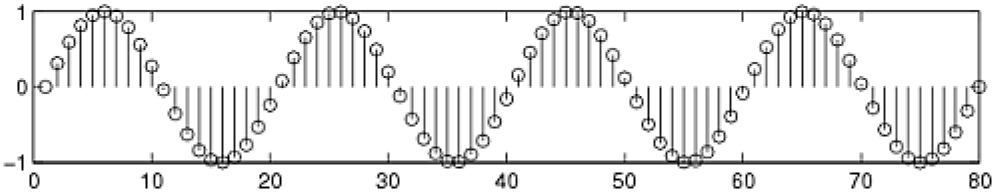
Odd part: $o(t) = \frac{1}{2}\left(f(t) - f(-t)\right)$



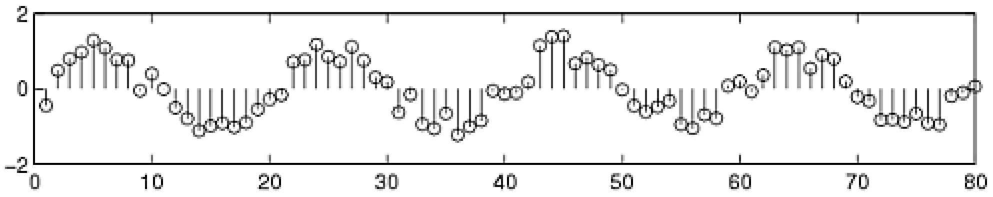
Check: $e(t) + o(t) = f(t)$

Deterministic vs. Random

A **deterministic signal** is a signal in which each value of the signal is fixed, being determined by a mathematical expression, rule, or table. On the other hand, the values of a **random signal** are not strictly defined, but are subject to some amount of variability.



Deterministic Signal



Random Signal

Example:
Consider the signal defined for all real t described by
Equation:

$$f(t) = \begin{cases} \sin(2\pi t)/t & t \geq 1 \\ 0 & t < 1 \end{cases}$$

This signal is continuous time, analog, aperiodic, infinite length, causal, neither even nor odd, and, by definition, deterministic.

Signal Classifications Summary

This module describes just some of the many ways in which signals can be classified. They can be continuous time or discrete time, analog or digital, periodic or aperiodic, finite or infinite, and deterministic or random. We can also divide them based on their causality and symmetry properties.

Linear Models

Finding an MVUB estimator is a very difficult task, in general. However, a large number of signal processing problems can be respresented by a **linear model** of the data.

Importance of Class of Linear Models

1. MVUB estimator within this class is immediately evident
2. Statistical performance analysis of linear models is very straightforward

General Form of Linear Model (LM)

$$\boldsymbol{x} = H\boldsymbol{\theta} + \boldsymbol{w}$$

where $\boldsymbol{x}$ is the **observation vector**, H is the known matrix (**observation** or **system matrix**), $\boldsymbol{\theta}$ is the unknown **parameter vector**, and $\boldsymbol{w}$ is the vector of White Guassian noise $\boldsymbol{w} \sim \mathcal{N}\left(0, \sigma^2 I\right)$.

Example:

$$\forall n, n \in \{1, \ldots, N\} : (x_n = A + B_n + w_n)$$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$$

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix}$$

$$\boldsymbol{\theta} = \begin{pmatrix} A \\ B \end{pmatrix}$$

$$H = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & N \end{pmatrix}$$

Probability Model for LM

$$\boldsymbol{x} = H\boldsymbol{\theta} + \boldsymbol{w}$$

$$x \sim \mathrm{p}\left(x \mid \theta\right) = \mathcal{N}\left(H\boldsymbol{\theta}, \sigma^2 I\right)$$

CRLB and NVUB Estimator

$\widehat{\boldsymbol{\theta}} = g(\boldsymbol{x})$ the MVUB estimator iff

$$\frac{\partial \log \mathrm{p}\left(x \mid \theta\right)}{\partial \boldsymbol{\theta}} = I(\boldsymbol{\theta})\left(g(\boldsymbol{x}) - \boldsymbol{\theta}\right)$$

In the case of the LM,

$$\frac{\partial \log \mathrm{p}\left(x \mid \theta\right)}{\partial \boldsymbol{\theta}} = \left(-\frac{1}{2\sigma^2}\right)\frac{\partial\left(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{x}^T H\boldsymbol{\theta} + \boldsymbol{\theta}^T H^T H\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}$$

Now using identities

$$\frac{\partial\left(\boldsymbol{b}^T\theta\right)}{\partial \boldsymbol{\theta}} = \boldsymbol{b}$$

$$\frac{\partial\left(\boldsymbol{\theta}^T A\boldsymbol{\theta}\right)}{\partial \theta} = 2A\boldsymbol{\theta}$$

for A symmetric.

We have

$$\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial \boldsymbol{\theta}} = \frac{1}{\sigma^2}\left(H^T\boldsymbol{x} - H^T H\boldsymbol{\theta}\right)$$

Assuming $H^T H$ is invertible

$$\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial \boldsymbol{\theta}} = \frac{H^T H}{\sigma^2}\left(\left(H^T H\right)^{-1} H^T\boldsymbol{x} - \boldsymbol{\theta}\right)$$

which leads to
Equation:

MVUB Estimator

$$\widehat{\boldsymbol{\theta}} = \left(H^T H\right)^{-1} H^T\boldsymbol{x}$$

Equation:

Fisher Information Matrix

$$I(\boldsymbol{\theta}) = \frac{H^T H}{\sigma^2}$$

$$\left\langle \widehat{\boldsymbol{\theta}}^2 \right\rangle = C_\theta = (I(\boldsymbol{\theta}))^{-1} = \sigma^2\left(H^T H\right)^{-1}$$

MVUB Estimator for the LM

If the observed data can be modeled as

$$\boldsymbol{x} = H\boldsymbol{\theta} + \boldsymbol{w}$$

where $\boldsymbol{w} \sim \mathcal{N}\left(0, \sigma^2 I\right)$ and H is invertible. Then, the MVUB estimator is

$$\widehat{\boldsymbol{\theta}} = \left(H^T H\right)^{-1} H^T \boldsymbol{x}$$

and the covariance of $\widehat{\boldsymbol{\theta}}$ is

$$C_\theta = \sigma^2 \left(H^T H\right)^{-1}$$

and $\widehat{\boldsymbol{\theta}}$ attains the CRLB.

Note: $\widehat{\boldsymbol{\theta}} \sim \mathscr{N}\left(\boldsymbol{\theta}, \sigma^2 \left(H^T H\right)^{-1}\right)$

Linear Model Examples

Example:
Curve Fitting
[missing_resource:]
Model:

$$\forall n, n \in \{1, \ldots, N\} : \left(x(t_n) = \theta_1 + \theta_2 t_n + \ldots + \theta_p {t_n}^{p-1} + w(t_n)\right)$$

where $\theta_1 + \theta_2 t_n + \ldots + \theta_p {t_n}^{p-1}$ is a $(p-1)^{\text{st}}$-order polynomial and $w(t_n) \sim \mathscr{N}\left(0, \sigma^2\right)$ idd. Therefore,

$$\boldsymbol{x} = H\boldsymbol{\theta} + \boldsymbol{w}$$

$$\boldsymbol{x} = \begin{pmatrix} x(t) \\ \vdots \\ x(t_n) \end{pmatrix}$$

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}$$

$$H = \begin{pmatrix} 1 & t_1 & \ldots & {t_1}^{p-1} \\ 1 & t_2 & \ldots & {t_2}^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_N & \ldots & {t_N}^{p-1} \end{pmatrix}$$

where H is the **Vandermonde matrix**. The MVUB estimator for $\boldsymbol{\theta}$ is

$$\hat{\theta} = \left(H^T H\right)^{-1} H^T \boldsymbol{x}$$

Example:
System Identification
[missing_resource:]

$$H(z) = \sum_{k=0}^{m-1} h[k] z^{-k}$$

$$\forall n, n \in \{0, \ldots, N-1\} : \left(x[n] = \sum_{k=0}^{m-1} h[k] u[n-k] + w[n]\right)$$

Where $w[n] \sim \mathcal{N}\left(0, \sigma^2\right)$ idd. Given x and u, estimate h.
In matrix form

$$\boldsymbol{x} = \begin{pmatrix} u[0] & 0 & \dots & 0 \\ u[1] & u[0] & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u[N-1] & u[N-2] & \dots & u[N-m] \end{pmatrix} \begin{pmatrix} h[0] \\ \vdots \\ \vdots \\ h[N-m] \end{pmatrix} + \boldsymbol{w}$$

where

$$\begin{pmatrix} u[0] & 0 & \dots & 0 \\ u[1] & u[0] & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u[N-1] & u[N-2] & \dots & u[N-m] \end{pmatrix} = H$$

and

$$\begin{pmatrix} h[0] \\ \vdots \\ \vdots \\ h[N-m] \end{pmatrix} = \boldsymbol{\theta}$$

Equation:

MVUB estimator

$$\widehat{\boldsymbol{\theta}} = \left(H^T H\right)^{-1} H^T \boldsymbol{x}$$

$$\left\langle \widehat{\boldsymbol{\theta}}^2 \right\rangle = \sigma^2 \left(H^T H\right)^{-1} = C_{\hat{\theta}}$$

An important question in system identification is how to choose the input $u[n]$ to "probe" the system most efficiently.

First note that

$$\sigma\left(\hat{\theta}_i\right)^2 = {e_i}^T C_{\hat{\theta}} e_i$$

where $e_i = (0\ldots010\ldots0)^T$. Also, since ${C_{\hat{\theta}}}^{-1}$ is symmetric positive definite, we can factor it by

$${C_{\hat{\theta}}}^{-1} = D^T D$$

where D is invertible.[footnote] Note that
Equation:

$$\left({e_i}^T D^T \left(D^T\right)^{-1} e_i\right)^2 = 1$$

The Schwarz inequality shows that [link] can become
Equation:

$$1 \leq \left({e_i}^T D^T D e_i\right) \left({e_i}^T D^{-1} \left(D^T\right)^{-1} e_i\right)$$

$$1 = \left({e_i}^T {C_{\hat{\theta}}}^{-1} e_i\right) \left({e_i}^T C_{\hat{\theta}} e_i\right)$$

which leads to

$$\sigma\left(\hat{\theta}_i\right)^2 \geq \frac{1}{{e_i}^T {C_{\hat{\theta}}}^{-1} e_i} = \frac{\sigma^2}{\left(H^T H\right)_{i,i}}$$

The minimum variance is achieved when equality is attained in [link]. This happens only if $\eta_1 = De_i$ is proportional to $\eta_2 = D^T e_i$. That is, $\eta_1 = C\eta_2$ for some constant C. Equivalently,

$$\forall i, i \in \{1, 2, \ldots, m\} : \left(D^T D e_i = c_i e_u\right)$$

$$D^T D = {C_{\hat{\theta}}}^{-1} = \frac{H^T H}{\sigma^2}$$

which leads to

$$\frac{H^T H}{\sigma^2} e_i = c_i e_i$$

Combining these equations in matrix form

$$H^T H = \sigma^2 \begin{pmatrix} c_1 & 0 & \dots & 0 \\ 0 & c_2 & \dots & 0 \\ 0 & 0 & \dots & c_m \end{pmatrix}$$

Therefore, in order to minimize the variance of the MVUB estimator, $u[n]$ should be chosen to make $H^T H$ diagonal.

Cholesky factorization

$$\forall i \wedge j, i \ \wedge \ j \in \{1, \dots, m\} : \left(\left(H^T H\right)_{i,j} = \sum_{n=1}^{N} u[n-i] u[n-j] \right)$$

For large N, this can be approximated to

$$\left(H^T H\right)_{i,j} \simeq \sum_{n=0}^{N-1-|i-j|} u[n] u[n+|i-j|]$$

using the autocorrelation of seq. $u[n]$.

Note: $u[n] = 0$ for $n < 0$ and $n > N-1$, letting limit of sum $-\infty, \infty$ gives approx.

These steps lead to

$$H^T H \simeq N \begin{pmatrix} r_{\mathrm{uu}}[0] & r_{\mathrm{uu}}[1] & \dots & r_{\mathrm{uu}}[m-1] \\ r_{\mathrm{uu}}[1] & r_{\mathrm{uu}}[0] & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ r_{\mathrm{uu}}[m-1] & r_{\mathrm{uu}}[m-2] & \dots & r_{\mathrm{uu}}[0] \end{pmatrix}$$

where $\begin{pmatrix} r_{\mathrm{uu}}[0] & r_{\mathrm{uu}}[1] & \dots & r_{\mathrm{uu}}[m-1] \\ r_{\mathrm{uu}}[1] & r_{\mathrm{uu}}[0] & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ r_{\mathrm{uu}}[m-1] & r_{\mathrm{uu}}[m-2] & \dots & r_{\mathrm{uu}}[0] \end{pmatrix}$ is the Toeplitz autocorrelation matrix and

$$r_{\mathrm{uu}}[n] = \frac{1}{N} \sum_{n=0}^{N-1-k} u[n]u[n+k]$$

For $H^T H$ to be diagonal, we require $r[k] = 0$, $k \neq 0$. This condition is approximately realized if we take $u[n]$ to be a **pseudorandom noise sequence** (PRN)[footnote]. Furthermore, the PRN sequence simplifies the estimator computation:

$$\hat{\boldsymbol{\theta}} = \left(H^T H\right)^{-1} H^T \boldsymbol{x}$$

$$\hat{\boldsymbol{\theta}} \simeq \frac{I}{N r_{\mathrm{uu}}[0]} H^T \boldsymbol{x}$$

which leads to

$$\widehat{h[i]} \simeq \frac{1}{N r_{\mathrm{uu}}[0]} \sum_{n=0}^{N-1-i} u[n-i]x[n]$$

where $\sum_{n=0}^{N-1-i} u[n-i]x[n] = N r_{\mathrm{ux}}[i]$. $r_{\mathrm{ux}}[i]$ is the cross-correlation between input and output sequences.

maximal length sequences

Hence, the approximate MVUB estimator for large N with a PRN input is

$$\forall i, i \in \{0, 1, \ldots, m-1\} : \left(\widehat{h[i]} = \frac{r_{\text{ux}}[i]}{r_{\text{uu}}[0]}\right)$$

$$r_{\text{ux}}[i] = \frac{1}{N}\sum_{n=0}^{N-1-i} u[n]x[n+i]$$

$$r_{\text{uu}}[0] = \frac{1}{N}\sum_{n=0}^{N-1} u^2[n]$$

CRLB for Signal in White Gaussian Noise

$$\forall n, n \in \{1, \ldots, N\} : (x_n = s_n(\theta) + w_n)$$

$$\mathrm{p}\,(x\,|\,\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\left(\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - s_n(\theta))^2\right)}$$

$$\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - s_n(\theta))\frac{\partial\, s_n(\theta)}{\partial\theta}$$

$$\frac{\partial^2 \log \mathrm{p}\,(x\,|\,\theta)}{\partial \text{msup}} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\left((x_n - s_n(\theta))\frac{\partial^2 s_n(\theta)}{\partial \text{msup}} - \left(\frac{\partial\, s_n(\theta)}{\partial\theta}\right)^2\right)$$

$$E\left[\frac{\partial^2 \log \mathrm{p}\,(x\,|\,\theta)}{\partial \text{msup}}\right] = -\left(\frac{1}{\sigma^2}\sum_{n=1}^{N}\left(\frac{\partial\, s_n(\theta)}{\partial\theta}\right)^2\right)$$

$$\sigma\left(\hat{\theta}\right)^2 \geq \frac{\sigma^2}{\sum_{n=1}^{N}\left(\frac{\partial\, s_n(\theta)}{\partial\theta}\right)^2}$$

Hypothesis Testing

Suppose you measure a collection of scalars $x_1, \ldots, x_N$. You believe the data is distributed in one of two ways. Your first model, call it H_0, postulates the data to be governed by the density $f_0(x)$ (some fixed density). Your second model, H_1, postulates a different density $f_1(x)$. These models, termed **hypotheses**, are denoted as follows:

$$H_0 : x_n \sim f_0(x), n = 1 \ldots N$$

$$H_1 : x_n \sim f_1(x), n = 1 \ldots N$$

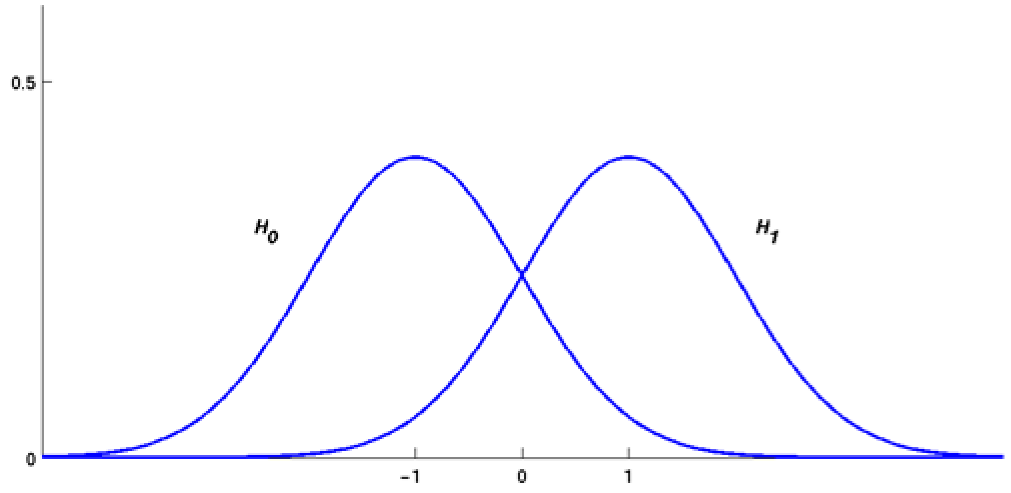
A **hypothesis test** is a rule that, given a measurement $\boldsymbol{x}$, makes a decision as to which hypothesis best "explains" the data.

Example:
Suppose you are confident that your data is normally distributed with variance 1, but you are uncertain about the sign of the mean. You might postulate

$$H_0 : x_n \sim \mathcal{N}(-1, 1)$$

$$H_1 : x_n \sim \mathcal{N}(1, 1)$$

These densities are depicted in [link].



Assuming each hypothesis is a priori equally likely, an intuitively appealing hypothesis test is to compute the sample mean $\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$, and choose H_0 if $\overline{x} \leq 0$, and H_1 if $\overline{x} > 0$. As we will see later, this test is in fact optimal under certain assumptions.

Generalizations and Nomenclature

The concepts introduced above can be extended in several ways. In what follows we provide more rigorous definitions, describe different kinds of hypothesis testing, and introduce terminology.

Data

In the most general setup, the observation is a collection $x_1, \ldots, x_N$ of random vectors. A common assumption, which facilitates analysis, is that the data are independent and identically distributed (IID). The random vectors may be continuous, discrete, or in some cases mixed. It is generally assumed that all of the data is available at once, although for some applications, such as Sequential Hypothesis Testing, the data is a never ending stream.

Binary Versus M-ary Tests

When there are two competing hypotheses, we refer to a **binary** hypothesis test. When the number of hypotheses is $M \geq 2$, we refer to an **M-ary** hypothesis test. Clearly, binary is a special case of M-ary, but binary tests are accorded a special status for certain reasons. These include their simplicity, their prevalence in applications, and theoretical results that do not carry over to the M-ary case.

Example:

Phase-Shift Keying

Suppose we wish to transmit a binary string of length r over a noisy communication channel. We assign each of the $M = 2^r$ possible bit sequences to a signal s^k, $k = \{1, \ldots, M\}$ where

$$s_n^k = \cos\left(2\pi f_0 n + \frac{2\pi (k-1)}{M}\right)$$

This symboling scheme is known as **phase-shift keying** (PSK). After transmitting a signal across the noisy channel, the receiver faces an M-ary hypothesis testing problem:

$$H_0 : \boldsymbol{x} = s^1 + \boldsymbol{w}$$

$$\vdots$$

$$H_M : \boldsymbol{x} = s^M + \boldsymbol{w}$$

where $\boldsymbol{w} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I\right)$.

In many binary hypothesis tests, one hypothesis represents the absence of a ceratin feature. In such cases, the hypothesis is usually labelled H_0 and called the **null** hypothesis. The other hypothesis is labelled H_1 and called the **alternative** hypothesis.

Example:

Waveform Detection

Consider the problem of detecting a known signal $\boldsymbol{s} = \left(s_1 \ldots s_N\right)^T$ in additive white Gaussian noise (AWGN). This scenario is common in sonar and radar systems. Denoting the data as $\boldsymbol{x} = \left(x_1 \ldots x_N\right)^T$, our hypothesis testing problem is

$$H_0 : \boldsymbol{x} = \boldsymbol{w}$$

$$H_1 : \boldsymbol{x} = \boldsymbol{s} + \boldsymbol{w}$$

where $\boldsymbol{w} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I\right)$. H_0 is the null hypothesis, corresponding to the absence of a signal.

Tests and Decision Regions

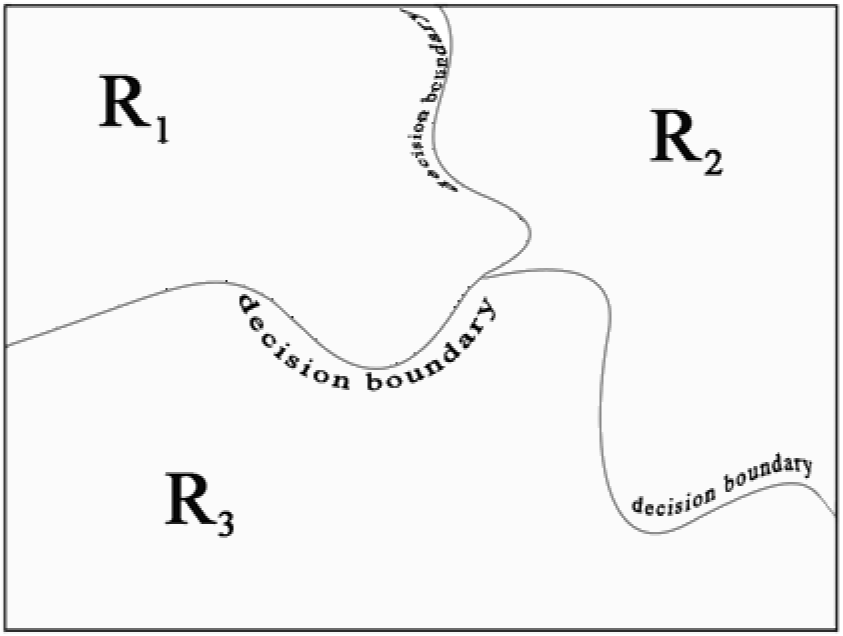
Consider the general hypothesis testing problem where we have N d-dimensional observations $x_1, \ldots, x_N$ and M hypotheses. If the data are real-valued, for example, then a hypothesis test is a mapping

$$\varphi : \left(\mathbb{R}^d\right)^N \to \{1, \ldots, M\}$$

For every possible realization of the input, the test outputs a hypothesis. The test φ partitions the input space into a disjoint collection $R_1, \ldots, R_M$, where

$$R_k = \{(x_1, \ldots, x_N) | \varphi(x_1, \ldots, x_N) = k\}$$

The sets R_k are called **decision regions**. The boundary between two decision regions is a **decision boundary**. [link] depicts these concepts when $d = 2$, $N = 1$, and $M = 3$.



Simple Versus Composite Hypotheses

If the distribution of the data under a certain hypothesis is fully known, we call it a **simple** hypothesis. All of the hypotheses in the examples above are simple. In many cases, however, we only know the distribution up to certain unknown parameters. For example, in a Gaussian noise model we may not know the variance of the noise. In this case, a hypothesis is said to be **composite**.

Example:
Consider the problem of detecting the signal

$$s_n = \cos(2\pi f_0 (n - k)) \forall n : (n = \{1, \ldots, N\})$$

where k is an unknown delay parameter. Then

$$H_0 : \boldsymbol{x} = \boldsymbol{w}$$

$$H_1 : \boldsymbol{x} = \boldsymbol{s} + \boldsymbol{w}$$

is a binary test of a simple hypothesis (H_0) versus a composite alternative. Here we are assuming $w_n \sim \mathcal{N}\left(0, \sigma^2\right)$, with σ^2 known.

Often a test involving a composite hypothesis has the form

$$H_0 : \boldsymbol{\theta} = \theta_0$$

$$H_1 : \boldsymbol{\theta} \neq \theta_0$$

where θ_0 is fixed. Such problems are called **two-sided** because the composite alternative "lies on both sides of H_0." When $\boldsymbol{\theta}$ is a scalar, the test

$$H_0 : \theta \leq \theta_0$$

$$H_1 : \theta > \theta_0$$

is called **one-sided**. Here, both hypotheses are composite.

Example:
Suppose a coin turns up heads with probability p. We want to assess whether the coin is fair ($p = \frac{1}{2}$). We toss the coin N times and record $x_1, \ldots, x_N$ ($x_n = 1$ means heads and $x_n = 0$ means tails). Then

$$H_0 : p = \frac{1}{2}$$

$$H_1 : p \neq \frac{1}{2}$$

is a binary test of a simple hypothesis (H_0) versus a composite alternative. This is also a two-sided test.

Errors and Probabilities

In binary hypothesis testing, assuming at least one of the two models does indeed correspond to reality, there are four possible scenarios:

- **Case 1** H_0 is true, and we declare H_0 to be true
- **Case 2** H_0 is true, but we declare H_1 to be true
- **Case 3** H_1 is true, and we declare H_1 to be true
- **Case 4** H_1 is true, but we declare H_0 to be true

In cases 2 and 4, errors occur. The names given to these errors depend on the area of application. In statistics, they are called **type I** and **type II errors** respectively, while in signal processing they are known as a **false alarm** or a **miss**.

Consider the general binary hypothesis testing problem

$$H_0 : \boldsymbol{x} \sim f_\theta(\boldsymbol{x}), \theta \in \Theta_0$$

$$H_1 : \boldsymbol{x} \sim f_\theta(\boldsymbol{x}), \theta \in \Theta_1$$

If H_0 is simple, that is, $\Theta_0 = \{\theta_0\}$, then the **size** (denoted α), also called the **false-alarm probability** (P_F), is defined to be

$$\alpha = P_F = \Pr[\theta_0, \mathrm{declare} H_1]$$

When Θ_0 is composite, we define

$$\alpha = P_F = \ \sup_{\theta \in \Theta_0} \left(\Pr[\theta, \mathrm{declare} H_1]\right)$$

For $\theta \in \Theta_1$, the **power** (denoted β), or **detection probability** (P_D), is defined to be

$$\beta = P_D = \Pr[\theta, \mathrm{declare} H_1]$$

The probability of a type II error, also called the **miss probability**, is

$$P_M = 1 - P_D$$

If H_1 is composite, then $\beta = \beta(\theta)$ is viewed as a function of θ.

Criteria in Hypothesis Testing

The design of a hypothesis test/detector often involves constructing the solution to an optimization problem. The optimality criteria used fall into two classes: Bayesian and frequent.

Representing the former approach is the Bayes Risk Criterion. Representing the latter is the Neyman-Pearson Criterion. These two approaches are developed at length in separate modules.

Statistics Versus Engineering Lingo

The following table, adapted from Kay, p.65, summarizes the different terminology for hypothesis testing from statistics and signal processing:

Statistics	**Signal Processing**
Hypothesis Test	Detector
Null Hypothesis	Noise Only Hypothesis
Alternate Hypothesis	Signal + Noise Hypothesis
Critical Region	Signal Present Decision Region
Type I Error	False Alarm
Type II Error	Miss
Size of Test (α)	Probability of False Alarm (P_F)
Power of Test (β)	Probability of Detection (P_D)

Criteria in Hypothesis Testing

The criterion used in the previous section - minimize the average cost of an incorrect decision - may seem to be a contrived way of quantifying decisions. Well, often it is. For example, the Bayesian decision rule depends explicitly on the a priori probabilities; a rational method of assigning values to these - either by experiment or through true knowledge of the relative likelihood of each model - may be unreasonable. In this section, we develop alternative decision rules that try to answer such objections. One essential point will emerge from these considerations: **the fundamental nature of the decision rule does not change with choice of optimization criterion**. Even criteria remote from error measures can result in the likelihood ratio test (see this problem). Such results do not occur often in signal processing and underline the likelihood ratio test's significance.

Maximum Probability of a Correct Decision

As only one model can describe any given set of data (the models are mutually exclusive), the probability of being correct P_c for distinguishing two models is given by

$$P_c = \Pr[\text{say } \mathscr{M}_0 \text{ when } \mathscr{M}_0 \text{ true}] + \Pr[\text{say } \mathscr{M}_1 \text{ when } \mathscr{M}_1 \text{ true}]$$

We wish to determine the optimum decision region placement Expressing the probability correct in terms of the likelihood functions $\mathrm{p}_{\boldsymbol{r}|\mathscr{M}_i}(\boldsymbol{r})$, the a priori probabilities, and the decision regions,

$$P_c = \int \pi_0 \, \mathrm{p}_{\boldsymbol{r}|\mathscr{M}_0}(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} + \int \pi_1 \, \mathrm{p}_{\boldsymbol{r}|\mathscr{M}_1}(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r}$$

We want to maximize P_c by selecting the decision regions $\mathfrak{R}_0$ and $\mathfrak{R}_0$. The probability correct is maximized by associating each value of $\boldsymbol{r}$ with the largest term in the expression for P_c. Decision region $\mathfrak{R}_0$, for example, is defined by the collection of values of $\boldsymbol{r}$ for which the first term is largest. As all of the quantities involved are non-negative, the decision rule maximizing the probability of a correct decision is

Note: Given $\boldsymbol{r}$, choose $\mathscr{M}_i$ for which the product $\pi_i \; \mathrm{p}_{\boldsymbol{r}|\mathscr{M}_i}(\boldsymbol{r})$ is largest.

Simple manipulations lead to the likelihood ratio test.

$$\frac{\mathrm{p}_{\boldsymbol{r}|\mathscr{M}_1}(\boldsymbol{r})}{\mathrm{p}_{\boldsymbol{r}|\mathscr{M}_0}(\boldsymbol{r})} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\pi_0}{\pi_1}$$

Note that if the Bayes' costs were chosen so that $C_{ii} = 0$ and $C_{ij} = C$, ($i \neq j$), we would have the same threshold as in the previous section.

To evaluate the quality of the decision rule, we usually compute the **probability of error** P_e rather than the probability of being correct. This quantity can be expressed in terms of the observations, the likelihood ratio, and the sufficient statistic.

Equation:

$$\begin{aligned} P_e &= \pi_0 \int \mathrm{p}_{\boldsymbol{r}|\mathscr{M}_0}(\boldsymbol{r}) \, \mathrm{d}\, \boldsymbol{r} + \pi_1 \int \mathrm{p}_{\boldsymbol{r}|\mathscr{M}_1}(\boldsymbol{r}) \, \mathrm{d}\, \boldsymbol{r} \\ &= \pi_0 \int \mathrm{p}_{\boldsymbol{\Lambda}|\mathscr{M}_0}(\boldsymbol{\Lambda}) \, \mathrm{d}\, \boldsymbol{\Lambda} + \pi_1 \int \mathrm{p}_{\boldsymbol{\Lambda}|\mathscr{M}_1}(\boldsymbol{\Lambda}) \, \mathrm{d}\, \boldsymbol{\Lambda} \\ &= \pi_0 \int \mathrm{p}_{\boldsymbol{\Upsilon}|\mathscr{M}_0}(\boldsymbol{\Upsilon}) \, \mathrm{d}\, \boldsymbol{\Upsilon} + \pi_1 \int \mathrm{p}_{\boldsymbol{\Upsilon}|\mathscr{M}_1}(\boldsymbol{\Upsilon}) \, \mathrm{d}\, \boldsymbol{\Upsilon} \end{aligned}$$

When the likelihood ratio is non-monotonic, the first expression is most difficult to evaluate. When monotonic, the middle expression proves the most difficult. Furthermore, these expressions point out that the likelihood ratio and the sufficient statistic can be considered a function of the observations $\boldsymbol{r}$; hence, they are random variables and have probability densities for each model. Another aspect of the resulting probability of error is that **no other decision rule can yield a lower probability of error**. This statement is obvious as we minimized the probability of error in deriving the likelihood ratio test. The point is that these expressions represent a lower bound on performance (as assessed by the probability of error). This probability will be non-zero if the conditional densities overlap over some range of values of $\boldsymbol{r}$, such as occurred in the previous example. In this region of overlap, the observed values are ambiguous: either model is consistent with the observations. Our "optimum" decision rule operates in such regions by selecting that model which is most likely (has the highest probability) of generating any particular value.

Neyman-Pearson Criterion

Situations occur frequently where assigning or measuring the a priori probabilities P_i is unreasonable. For example, just what is the a priori probability of a supernova occurring in any particular region of the sky? We clearly need a model evaluation procedure which can function without a priori probabilities. This kind of test results when the so-called Neyman-Pearson criterion is used to derive the decision rule. The ideas behind and decision rules derived with the Neyman-Pearson criterion (Neyman and Pearson) will serve us well in sequel; their result is important!

Using nomenclature from radar, where model $\mathscr{M}_1$ represents the presence of a target and $\mathscr{M}_0$ its absence, the various types of correct and incorrect decisions have the following names (Woodward, pp. 127-129).[footnote]

- **Detection** we say it's there when it is; $P_D = \Pr\left(\text{say } \ \mathscr{M}_1 | \mathscr{M}_1 \ \text{ true}\right)$
- **False-alarm** we say it's there when it's not; $P_F = \Pr\left(\text{say } \ \mathscr{M}_1 | \mathscr{M}_0 \ \text{ true}\right)$
- **Miss** we say it's not there when it is; $P_M = \Pr\left(\text{say } \ \mathscr{M}_0 | \mathscr{M}_1 \ \text{ true}\right)$

The remaining probability $\Pr[\text{say } \ \mathscr{M}_0 \mid \mathscr{M}_0 \ \text{ true}]$ has historically been left nameless and equals $1 - P_F$. We should also note that the detection and miss probabilities are related by $P_M = 1 - P_D$. As these are conditional probabilities, they do not depend on the a priori probabilities and the two probabilities P_F and P_D characterize the errors when **any** decision rule is used.
In hypothesis testing, a false-alarm is known as a **type I error** and a miss a **type II error**.

These two probabilities are related to each other in an interesting way. Expressing these quantities in terms of the decision regions and the likelihood functions, we have

$$P_F = \int \mathrm{p}_{\boldsymbol{r} | \mathscr{M}_0}(\boldsymbol{r})\,\mathrm{d}\,\boldsymbol{r}$$

$$P_D = \int \mathrm{p}_{\boldsymbol{r} | \mathscr{M}_1}(\boldsymbol{r})\,\mathrm{d}\,\boldsymbol{r}$$

As the region $\mathfrak{R}_1$ shrinks, **both** of these probabilities tend toward zero; as $\mathfrak{R}_1$ expands to engulf the entire range of observation values, they both tend toward unity. This rather direct relationship between P_D and P_F does not mean that they equal each other; in most cases, as $\mathfrak{R}_1$ expands, P_D increases more rapidly than P_F (we had better be right more often than we are wrong!). However, the "ultimate" situation where a rule is always right and never wrong ($P_D = 1$, $P_F = 0$) cannot occur when the conditional distributions overlap. Thus, to increase the detection probability we must also allow the false-alarm probability to increase. This behavior represents the fundamental tradeoff in hypothesis testing and detection theory.

One can attempt to impose a performance criterion that depends only on these probabilities with the consequent decision rule not depending on the a priori probabilities. The Neyman-Pearson criterion assumes that the false-alarm probability is constrained to be less than or equal to a specified value α while we attempt to maximize the detection probability P_D.

$$\forall P_F, P_F \leq \alpha : \left(\max_{\mathfrak{R}_1} \{\mathfrak{R}_1, P_D\}\right)$$

A subtlety of the succeeding solution is that the underlying probability distribution functions may not be continuous, with the result that P_F can never equal the constraining value α. Furthermore, an (unlikely) possibility is that the optimum value for the false-alarm probability is somewhat less than the criterion value. Assume, therefore, that we rephrase the optimization problem by requiring that the false-alarm probability equal a value α' that is less than or equal to α.

This optimization problem can be solved using Lagrange multipliers (see Constrained Optimization); we seek to find the decision rule that maximizes

$$F = P_D + \lambda \left(P_F - \alpha'\right)$$

where λ is the Lagrange multiplier. This optimization technique amounts to finding the decision rule that maximizes F, then finding the value of the multiplier that allows the criterion to be satisfied. As is usual in the derivation of optimum decision rules, we maximize these quantities with respect to the decision regions. Expressing P_D and P_F in terms of them, we have

Equation:

$$\begin{aligned} F &= \int \mathrm{p}_{\boldsymbol{r}|\mathscr{M}_1}(\boldsymbol{r})\,\mathrm{d}\,\boldsymbol{r} + \lambda \left(\int \mathrm{p}_{\boldsymbol{r}|\mathscr{M}_0}(\boldsymbol{r})\,\mathrm{d}\,\boldsymbol{r} - \alpha'\right) \\ &= -(\lambda\alpha') + \int \mathrm{p}_{\boldsymbol{r}|\mathscr{M}_1}(\boldsymbol{r}) + \lambda\,\mathrm{p}_{\boldsymbol{r}|\mathscr{M}_0}(\boldsymbol{r})\,\mathrm{d}\,\boldsymbol{r} \end{aligned}$$

To maximize this quantity with respect to $\mathfrak{R}_1$, we need only to integrate over those regions of $\boldsymbol{r}$ where the integrand is positive. The region $\mathfrak{R}_1$ thus corresponds to those values of $\boldsymbol{r}$ where $\mathrm{p}_{\boldsymbol{r}|\mathscr{M}_1}(\boldsymbol{r}) > -\left(\lambda\,\mathrm{p}_{\boldsymbol{r}|\mathscr{M}_0}(\boldsymbol{r})\right)$ and the resulting decision rule is

$$\frac{\mathrm{p}_{\boldsymbol{r}|\mathscr{M}_1}(\boldsymbol{r})}{\mathrm{p}_{\boldsymbol{r}|\mathscr{M}_0}(\boldsymbol{r})} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} (-\lambda)$$

The ubiquitous likelihood ratio test again appears; it **is** indeed the fundamental quantity in hypothesis testing. Using the logarithm of the likelihood ratio or the sufficient statistic, this result can be expressed as either

$$\ln(\mathit{\Lambda}(\boldsymbol{r})) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln(-\lambda)$$

or

$$\mathit{\Upsilon}(\boldsymbol{r}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma$$

We have not as yet found a value for the threshold. The false-alarm probability can be expressed in terms of the Neyman-Pearson threshold in two (useful) ways.

Equation:

$$\begin{aligned} P_F &= \int_{-\lambda}^{\infty} \mathrm{p}_{\boldsymbol{\Lambda} | \mathscr{M}_0} \left(\boldsymbol{\Lambda}\right) \, \mathrm{d} \, \Lambda \\ &= \int_{\gamma}^{\infty} \mathrm{p}_{\boldsymbol{\Upsilon} | \mathscr{M}_0} \left(\boldsymbol{\Upsilon}\right) \, \mathrm{d} \, \Upsilon \end{aligned}$$

One of these implicit equations must be solved for the threshold by setting P_F equal to α'. The selection of which to use is usually based on pragmatic considerations: the easiest to compute. From the previous discussion of the relationship between the detection and false-alarm probabilities, we find that to maximize P_D we must allow α' to be as large as possible while remaining less than α. Thus, we want to find the **smallest** value of $-\lambda$ (note the minus sign) consistent with the constraint. Computation of the threshold is problem-dependent, but a solution always exists.

Example:

An important application of the likelihood ratio test occurs when $\boldsymbol{r}$ is a Gaussian random vector for each model. Suppose the models correspond to Gaussian random vectors having different mean values but sharing the same identity covariance.

- $\mathscr{M}_0$: $\boldsymbol{r} \sim \mathscr{N}\left(0, \sigma^2 I\right)$
- $\mathscr{M}_1$: $\boldsymbol{r} \sim \mathscr{N}\left(\boldsymbol{m}, \sigma^2 I\right)$

Thus, $\boldsymbol{r}$ is of dimension L and has statistically independent, equal variance components. The vector of means $\boldsymbol{m} = \left(m_0 \ldots m_{L-1}\right)^T$ distinguishes the two models. The likelihood functions associated this problem are

$$\mathrm{p}_{\boldsymbol{r} | \mathscr{M}_0} \left(\boldsymbol{r}\right) = \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(1/2\left(\frac{r_l}{\sigma}\right)^2\right)}$$

$$\mathrm{p}_{\boldsymbol{r} | \mathscr{M}_1} \left(\boldsymbol{r}\right) = \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(1/2\left(\frac{r_l - m_l}{\sigma}\right)^2\right)}$$

The likelihood ratio $\Lambda(\boldsymbol{r})$ becomes

$$\Lambda(\boldsymbol{r}) = \frac{\prod_{l=0}^{L-1} e^{-\left(1/2\left(\frac{r_l-m_l}{\sigma}\right)^2\right)}}{\prod_{l=0}^{L-1} e^{-\left(1/2\left(\frac{r_l}{\sigma}\right)^2\right)}}$$

This expression for the likelihood ratio is complicated. In the Gaussian case (and many others), we use the logarithm the reduce the complexity of the likelihood ratio and form a sufficient statistic.

Equation:

$$\begin{aligned} \ln(\Lambda(\boldsymbol{r})) &= \sum_{l=0}^{L-1} -1/2\frac{(r_l-m_l)^2}{\sigma^2} + 1/2\frac{{r_l}^2}{\sigma^2} \\ &= \frac{1}{\sigma^2}\sum_{l=0}^{L-1} m_l r_l - \frac{1}{2\sigma^2}\sum_{l=0}^{L-1} {m_l}^2 \end{aligned}$$

The likelihood ratio test then has the much simpler, but equivalent form

$$\sum_{l=0}^{L-1} (m_l r_l) \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} (\sigma^2 \ln(\eta)) + 1/2\sum_{l=0}^{L-1} {m_l}^2$$

To focus on the model evaluation aspects of this problem, let's assume means be equal to a positive constant: $m_l = m$ (0).[footnote]

$$\sum_{l=0}^{L-1} r_l \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} \left(\frac{\sigma^2}{m}\ln(\eta)\right) + \frac{Lm}{2}$$

Note that all that need be known about the observations $\{r_l\}$ is their sum. This quantity is the sufficient statistic for the Gaussian problem: $\Upsilon(\boldsymbol{r}) = \sum r_l$ and $\gamma = \sigma^2 \ln\left(\frac{\eta}{m}\right) + \frac{Lm}{2}$.

Why did the authors assume that the mean was positive? What would happen if it were negative?

When trying to compute the probability of error or the threshold in the Neyman-Pearson criterion, we must find the conditional probability density of one of the decision statistics: the likelihood ratio, the log-likelihood, or the sufficient statistic. The log-likelihood and the sufficient statistic are quite similar in this problem, but clearly we should use the latter. One practical property of the sufficient statistic is that it usually simplifies computations. For this Gaussian example, the sufficient statistic is a Gaussian random variable under each model.

- $\mathscr{M}_0$: $\boldsymbol{\Upsilon}(\mathbf{r}) \sim \mathscr{N}\left(0, L\sigma^2\right)$
- $\mathscr{M}_1$: $\boldsymbol{\Upsilon}(\mathbf{r}) \sim \mathscr{N}\left(Lm, L\sigma^2\right)$

To find the probability of error from [link], we must evaluate the area under a Gaussian probability density function. These integrals are succinctly expressed in terms of $Q(x)$, which denotes the probability that a unit-variance, zero-mean Gaussian random variable exceeds x (see Probability and Stochastic Processes). As $1 - Q(x) = Q(-x)$, the probability of error can be written as

$$P_e = \pi_1 Q\left(\frac{Lm - \gamma}{\sqrt{L}\sigma}\right) + \pi_0 Q\left(\frac{\gamma}{\sqrt{L}\sigma}\right)$$

An interesting special case occurs when $\pi_0 = 1/2 = \pi_1$. In this case, $\gamma = \frac{Lm}{2}$ and the probability of error becomes

$$P_e = Q\left(\frac{\sqrt{L}m}{2\sigma}\right)$$

As $Q(\cdot)$ is a monotonically decreasing function, the probability of error decreases with increasing values of the ratio $\frac{\sqrt{L}m}{2\sigma}$. However, as shown in this figure, $Q(\cdot)$ decreases in a nonlinear fashion. Thus, increasing m by a factor of two may decrease the probability of error by a larger **or** a smaller factor; the amount of change depends on the initial value of the ratio.
To find the threshold for the Neyman-Pearson test from the expressions given on [link], we need the area under a Gaussian density.
Equation:

$$\begin{aligned} P_F &= Q\left(\frac{\gamma}{\sqrt{L\sigma^2}}\right) \\ &= \alpha' \end{aligned}$$

As $Q(\cdot)$ is a monotonic and continuous function, we can now set α' equal to the criterion value α with the result

$$\gamma = \sqrt{L}\sigma Q^{-1}(\alpha)$$

where $Q^{-1}(\cdot)$ denotes the inverse function of $Q(\cdot)$. The solution of this equation cannot be performed analytically as no closed form expression exists for $Q(\cdot)$ (much less its inverse function); the criterion value must be found from tables or numerical routines. Because Gaussian problems arise frequently, the [link] accompanying table provides numeric values for this quantity at the decade points.

x	$Q^{-1}(x)$
10^{-1}	1.281
10^{-2}	2.396
10^{-3}	3.090
10^{-4}	3.719
10^{-5}	4.265
10^{-6}	4.754

The table displays interesting values for $Q^{-1}(\cdot)$ that can be used to determine thresholds in the Neyman-Pearson variant of the likelihood ratio test. Note how little the inverse function changes for decade changes in its argument; $Q(\cdot)$ is indeed **very** nonlinear.

The detection probability is given by

$$P_D = Q\left(Q^{-1}(\alpha) - \frac{\sqrt{L}m}{\sigma}\right)$$

The Bayes Risk Criterion in Hypothesis Testing

The design of a hypothesis test/detector often involves constructing the solution to an optimization problem. The optimality criteria used fall into two classes: Bayesian and frequent.

In the Bayesian setup, it is assumed that the a priori probability of each hypothesis occuring (π_i) is known. A cost C_{ij} is assigned to each possible outcome:

$$C_{\mathrm{ij}} = \Pr[\mathrm{say} H_i \mathrm{when} H_j \mathrm{true}]$$

The optimal test/detector is the one that minimizes the Bayes risk, which is defined to be the expected cost of an experiment:

$$C = \sum_{i,j} C_{\mathrm{ij}} \pi_i \Pr[\mathrm{say} H_i \mathrm{when} H_j \mathrm{true}]$$

In the event that we have a binary problem, and both hypotheses are simple, the decision rule that minimizes the Bayes risk can be constructed explicitly. Let us assume that the data is continuous (i.e., it has a density) under each hypothesis:

$$H_0 : \boldsymbol{x} \sim f_0(\boldsymbol{x})$$

$$H_1 : \boldsymbol{x} \sim f_1(\boldsymbol{x})$$

Let R_0 and R_1 denote the decision regions corresponding to the optimal test. Clearly, the optimal test is specified once we specify R_0 and $R_1 = {R_0}'$.

The Bayes risk may be written
Equation:

$$\begin{aligned} \bar{C} &= \sum_{\{i,j\}=0}^{1} C_{ij}\pi_i \int f_j(\boldsymbol{x}) \, \mathrm{d}\, \boldsymbol{x} \\ &= \int C_{00}\pi_0 f_0(\boldsymbol{x}) + C_{01}\pi_1 f_1(\boldsymbol{x}) \, \mathrm{d}\, \boldsymbol{x} + \int C_{10}\pi_0 f_0(\boldsymbol{x}) + C_{11}\pi_1 f_1(\boldsymbol{x}) \, \mathrm{d}\, \boldsymbol{x} \end{aligned}$$

Recall that R_0 and R_1 **partition** the input space: they are disjoint and their union is the full input space. Thus, every possible input $\boldsymbol{x}$ belongs to precisely one of these regions. In order to minimize the Bayes risk, a measurement $\boldsymbol{x}$ should belong to the decision region R_i for which the corresponding integrand in the preceding equation is smaller. Therefore, the Bayes risk is minimized by assigning $\boldsymbol{x}$ to R_0 whenever

$$\pi_0 C_{00} f_0(\boldsymbol{x}) + \pi_1 C_{01} f_1(\boldsymbol{x}) < \pi_0 C_{10} f_0(\boldsymbol{x}) + \pi_1 C_{11} f_1(\boldsymbol{x})$$

and assigning $\boldsymbol{x}$ to R_1 whenever this inequality is reversed. The resulting rule may be expressed concisely as

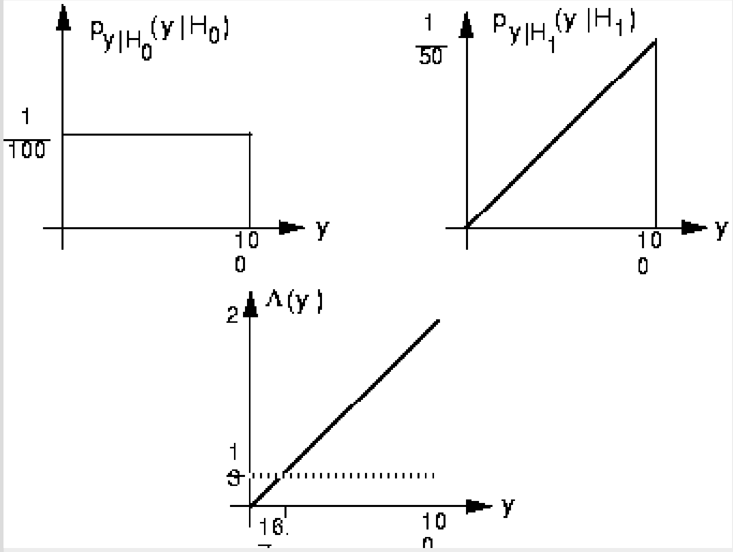
$$\Lambda(\boldsymbol{x}) \equiv \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\pi_0 \left(C_{10} - C_{00}\right)}{\pi_1 \left(C_{01} - C_{11}\right)} \equiv \eta$$

Here, $\Lambda(\boldsymbol{x})$ is called the **likelihood ratio**, η is called the threshold, and the overall decision rule is called the Likelihood Ratio Test (LRT). The expression on the right is called a **threshold**.

Example:
An instructor in a course in detection theory wants to determine if a particular student studied for his last test. The observed quantity is the student's grade, which we denote by $\boldsymbol{r}$. Failure may not indicate studiousness: conscientious students may fail the test. Define the models as

- $\mathscr{M}_0$: did not study
- $\mathscr{M}_1$: did study

The conditional densities of the grade are shown in [link].



Conditional densities for the grade distributions assuming that a student did not study ($\mathscr{M}_0$) or did ($\mathscr{M}_1$) are shown in the top row. The lower portion depicts the likelihood ratio formed from these densities.

Based on knowledge of student behavior, the instructor assigns a priori probabilities of $\pi_0 = \frac{1}{4}$ and $\pi_1 = \frac{3}{4}$. The costs C_{ij} are chosen to reflect the instructor's sensitivity to student feelings: $C_{01} = 1 = C_{10}$ (an erroneous decision either way is given the same cost) and $C_{00} = 0 = C_{11}$. The likelihood ratio is plotted in [link] and the threshold value η, which is computed from the a priori probabilities and the costs to be $\frac{1}{3}$, is indicated. The calculations of this comparison can be simplified in an obvious way.

$$\frac{r}{50} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{1}{3}$$

or

$$r \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{50}{3} = 16.7$$

The multiplication by the factor of 50 is a simple illustration of the reduction of the likelihood ratio to a sufficient statistic. Based on the assigned costs and a priori probabilities, the optimum decision rule says the instructor must assume that the student did not study if the student's grade is less than 16.7; if greater, the student is assumed to have studied despite receiving an abysmally low grade such as 20. Note that as the densities given by each model overlap entirely: the possibility of making the wrong interpretation **always** haunts the instructor. However, no other procedure will be better!

A special case of the minimum Bayes risk rule, the minimum probability of error rule, is used extensively in practice, and is discussed at length in another module.

Problems

Exercise:

Problem:

Denote $\alpha = \Pr[\mathrm{declare} H_1 \mathrm{when} H_0 \mathrm{true}]$ and

$\beta = \Pr[\mathrm{declare} H_1 \mathrm{when} H_1 \mathrm{true}]$. Express the Bayes risk $\bar{C}$ in terms of α and β, C_{ij}, and π_i. Argue that the optimal decision rule is not altered by setting $C_{00} = C_{11} = 0$.

Exercise:

Problem:

Suppose we observe $\boldsymbol{x}$ such that $\Lambda(\boldsymbol{x}) = \eta$. Argue that it doesn't matter whether we assign $\boldsymbol{x}$ to R_0 or R_1.

Minimum Probability of Error Decision Rule

Consider the binary hypothesis test

$$\mathscr{H}_0 : \boldsymbol{x} \sim f_0(\boldsymbol{x})$$

$$\mathscr{H}_1 : \boldsymbol{x} \sim f_1(\boldsymbol{x})$$

Let π_i, denote the a priori probability of hypothesis $\mathscr{H}_i$. Suppose our decision rule declares "$\mathscr{H}_0$ is the true model" when $\boldsymbol{x} \in R_0$, and it selects $\mathscr{H}_1$ when $\boldsymbol{x} \in R_1$, where $R_1 = {R_0}'$. The probability of making an error, denoted P_e, is

Equation:

$$\begin{aligned} P_e &= \Pr[\text{declare}\mathscr{H}_0\text{and}\mathscr{H}_1\text{true}] + \Pr[\text{declare}\mathscr{H}_1\text{and}\mathscr{H}_0\text{true}] \\ &= \Pr[\mathscr{H}_1]\Pr[\mathscr{H}_0 \mid \mathscr{H}_1] + \Pr[\mathscr{H}_0]\Pr[\mathscr{H}_1 \mid \mathscr{H}_0] \\ &= \int \pi_1 f_1(\boldsymbol{x})\,\mathrm{d}\,\boldsymbol{x} + \int \pi_0 f_0(\boldsymbol{x})\,\mathrm{d}\,\boldsymbol{x} \end{aligned}$$

In this module, we study the minimum probability of error decision rule, which selects R_0 and R_1 so as to minimize the above expression.

Since an observation $\boldsymbol{x}$ falls into one and only one of the decision regions R_i, in order to minimize P_e, we assign $\boldsymbol{x}$ to the region for which the corresponding integrand in [link] is smaller. Thus, we select $\boldsymbol{x} \in R_0$ if $\pi_1 f_1(\boldsymbol{x}) < \pi_0 f_0(\boldsymbol{x})$, and $\boldsymbol{x} \in R_1$ if the inequality is reversed. This decision rule may be summarized concisely as

$$\Lambda(\boldsymbol{x}) \equiv \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \frac{\pi_0}{\pi_1} \equiv \eta$$

Here, $\Lambda(\boldsymbol{x})$ is called the **likelihood ratio**, η is called a **threshold**, and the overall decision rule is called the Likelihood Ratio Test.

Example:

Normal with Common Variance, Uncommon Means

Consider the binary hypothesis test of a scalar x

$$\mathscr{H}_0 : x \sim \mathscr{N}\left(0, \sigma^2\right)$$

$$\mathscr{H}_1 : x \sim \mathscr{N}\left(\mu, \sigma^2\right)$$

where μ and σ^2 are known, positive quantities. Suppose we observe a single measurement x. The likelihood ratio is **Equation:**

$$\begin{aligned} \Lambda(x) &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}} \\ &= e^{\frac{1}{\sigma^2}\left(\mu x - \frac{\mu^2}{2}\right)} \end{aligned}$$

and so the minimum probability of error decision rule is

$$e^{\frac{1}{\sigma^2}\left(\mu x - \frac{\mu^2}{2}\right)} \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \frac{\pi_0}{\pi_1} = \eta$$

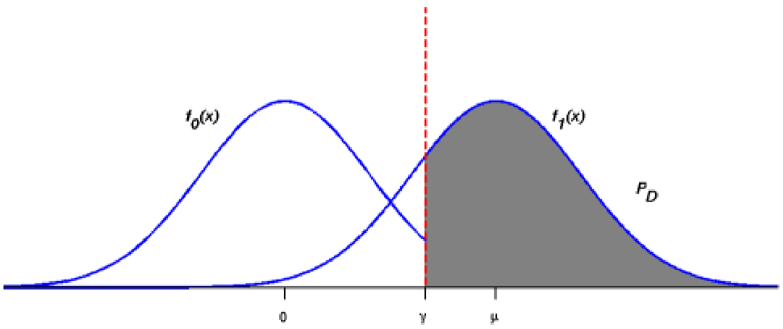
The expression for $\Lambda(x)$ is somewhat complicated. By applying a sequence of monotonically increasing functions to both sides, we can obtain a simplified expression for the optimal decision rule without changing the rule. In this example, we apply the natural logarithm and rearrange terms to arrive at

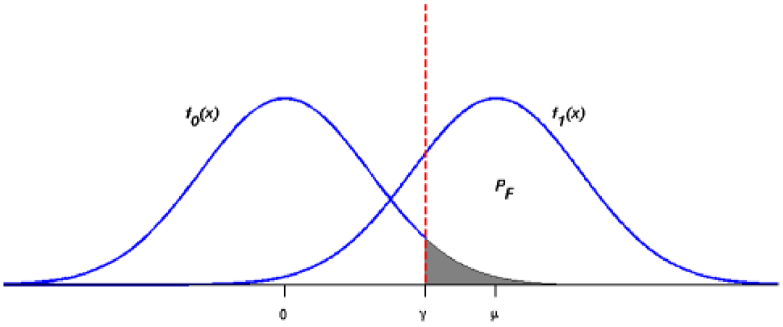
$$x \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \frac{\sigma^2}{\mu} \ln(\eta) + \frac{\mu}{2} \equiv \gamma$$

Here we have used the assumption $\mu > 0$. If $\mu < 0$, then dividing by μ would reverse the inequalities.

This form of the decision rule is much simpler: we just compare the observed value x to a threshold γ. [link] depicts the two candidate densities and a possible value of γ. If each hypothesis is a priori equally likely ($\pi_0 = \pi_1 = \frac{1}{2}$), then $\gamma = \frac{\mu}{2}$. [link] illustrates the case where $\pi_0 > \pi_1$ ($\gamma > \frac{\mu}{2}$).

The two candidate densities, and a threshold corresponding to $\pi_0 > \pi_1$



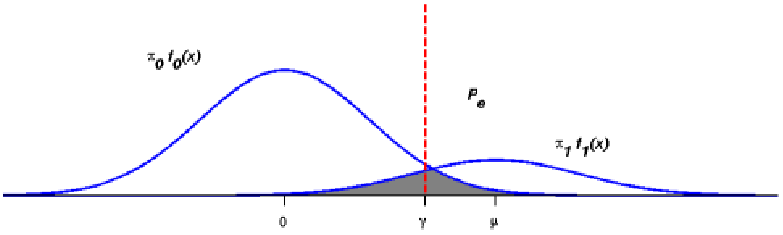


If we plot the two densities so that each is weighted by its a priori probability of occuring, the two curves will intersect at the threshold γ (see [link]). (Can you explain why this is? Think back to our derivation of the LRT). This plot also offers a way to visualize the probability of error. Recall

Equation:

$$\begin{aligned} P_e &= \int \pi_1 f_1(x) \,\mathrm{d}\, x + \int \pi_0 f_0(x) \,\mathrm{d}\, x \\ &= \int \pi_1 f_1(x) \,\mathrm{d}\, x + \int \pi_0 f_0(x) \,\mathrm{d}\, x \\ &= \pi_1 P_M + \pi_0 P_F \end{aligned}$$

where P_M and P_F denote the miss and false alarm probabilities, respectively. These quantities are depicted in [link].



The candidate densities weighted by their a priori probabilities. The shaded region is the probability of error for the optimal decision rule.

We can express P_M and P_F in terms of the Q-function as

$$P_e = \pi_1 Q\left(\frac{\mu - \gamma}{\sigma}\right) + \pi_0 Q\left(\frac{\gamma}{\sigma}\right)$$

When $\pi_0 = \pi_1 = \frac{1}{2}$, we have $\gamma = \frac{\mu}{2}$, and the error probability is

$$P_e = Q\left(\frac{\mu}{2\sigma}\right)$$

Since $Q(x)$ is monotonically decreasing, this says that the "difficulty" of the detection problem decreases with decreasing σ and increasing μ.

In the preceding example, computation of the probability of error involved a one-dimensional integral. If we had multiple observations, or vector-valued data, generalizing this procedure would involve multi-dimensional integrals over potentially complicated decision regions. Fortunately, in many cases, we can avoid this problem through the use of sufficient statistics.

Example:
Suppose we have the same test as in the previous example, but now we have N independent observations:

$$\mathscr{H}_0 : x_n \sim \mathscr{N}\left(0, \sigma^2\right), n = 1, \ldots, N$$

$$\mathscr{H}_1 : x_n \sim \mathscr{N}\left(\mu, \sigma^2\right), n = 1, \ldots, N$$

where $\mu > 0$ and $\sigma^2 > 0$ and both are known. The likelihood ratio is
Equation:

$$\begin{aligned} \Lambda(\boldsymbol{x}) &= \frac{\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}}{\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_n{}^2}{2\sigma^2}}} \\ &= \frac{e^{\frac{-1}{2\sigma^2} \sum_{n=1}^{N} (x_n-\mu)^2}}{e^{\frac{-1}{2\sigma^2} \sum_{n=1}^{N} x_n{}^2}} \\ &= e^{\frac{1}{2\sigma^2} \sum_{n=1}^{N} 2x_n\mu-\mu^2} \\ &= e^{\frac{1}{\sigma^2} \left(\mu \sum_{n=1}^{N} x_n - \frac{N\mu^2}{2}\right)} \end{aligned}$$

As in the previous example, we may apply the natural logarithm and rearrange terms to obtain an equivalent form of the LRT:

$$t \equiv \sum_{n=1}^{N} x_n \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \frac{\sigma^2}{\mu} \ln(\eta) + \frac{N\mu}{2} \equiv \gamma$$

The scalar quantity t is a sufficient statistic for the mean. In order to evaluate the probability of error without resorting to a multi-dimensional integral, we can express P_e in terms of t as

$$P_e = \pi_1 \Pr[t < \gamma \mid \mathscr{H}_1 \text{true}] + \pi_0 \Pr[t > \gamma \mid \mathscr{H}_0 \text{true}]$$

Now t is a linear combination of normal variates, so it is itself normal. In particular, we have $t = \boldsymbol{Ax}$, where $(1 \quad \dots \quad 1)$ is an N-dimensional row vector of 1's, and $\boldsymbol{x}$ is multivariate normal with mean $\mathbf{0}$ or $\boldsymbol{\mu} = (\mu \ldots \mu)^T$, and covariance $\sigma^2 I$. Thus we have

$$t|\mathscr{H}_0 \sim \mathcal{N}\left(\boldsymbol{A}\mathbf{0}, \boldsymbol{A}\sigma^2 I \boldsymbol{A}^T\right) = \mathcal{N}\left(0, N\sigma^2\right)$$

$$t|\mathscr{H}_1 \sim \mathcal{N}\left(\boldsymbol{A\mu}, \boldsymbol{A}\sigma^2 I \boldsymbol{A}^T\right) = \mathcal{N}\left(N\mu, N\sigma^2\right)$$

Therefore, we may write P_e in terms of the Q-function as

$$P_e = \pi_1 Q\left(\frac{N\mu - \gamma}{\sqrt{N}\sigma}\right) + \pi_0 Q\left(\frac{\gamma}{\sqrt{N}\sigma}\right)$$

In the special case $\pi_0 = \pi_1 = \frac{1}{2}$,

$$P_e = Q\left(\frac{\sqrt{N}\mu}{\sigma}\right)$$

Since Q is monotonically decreasing, this result provides mathematical support for something that is intuitively obvious: The performance of our decision rule improves with increasing N and μ, and decreasing σ.

Note: In the context of signal processing, the foregoing problem may be viewed as the problem of detecting a constant (DC) signal in additive white Gaussian noise:

$$\mathscr{H}_0 : x_n = w_n, n = 1, \ldots, N$$

$$\mathscr{H}_1 : x_n = A + w_n, n = 1, \ldots, N$$

where A is a known, fixed amplitude, and $w_n \sim \mathscr{N}\left(0, \sigma^2\right)$. Here A corresponds to the mean μ in the example.

The next example explores the minimum probability of error decision rule in a **discrete** setting.

Example:

Repetition Code

Suppose we have a friend who is trying to transmit a bit (0 or 1) to us over a noisy channel. The channel causes an error in the transmission (that is, the bit is flipped) with probability p, where $0 \leq p < \frac{1}{2}$, and p is known. In order to increase the chance of a successful transmission, our friend sends the same bit N times. Assume the N transmissions are statistically independent. Under these assumptions, the bits you receive are Bernoulli random variables: $x_n \sim \text{Bernoulli}\,(\theta)$. We are faced with the following hypothesis test:

$\mathscr{H}_0$	$\theta = p$	0 sent

$\mathscr{H}_0$	$\theta = p$	0 sent
$\mathscr{H}_1$	$\theta = 1 - p$	1 sent

We decide to decode the received sequence $\boldsymbol{x} = (x_1 \ldots x_N)^T$ by minimizing the probability of error. The likelihood ratio is

Equation:

$$\begin{aligned} \Lambda(\boldsymbol{x}) &= \frac{\prod_{n=1}^{N} (1-p)^{x_n} p^{1-x_n}}{\prod_{n=1}^{N} p^{x_n} (1-p)^{1-x_n}} \\ &= \frac{(1-p)^k p^{N-k}}{p^k (1-p)^{N-k}} \\ &= \left(\frac{1-p}{p}\right)^{2k-N} \end{aligned}$$

where $k = \sum_{n=1}^{N} x_n$ is the number of 1s received.

Note: k is a sufficient statistic for θ.

The LRT is

$$\left(\frac{1-p}{p}\right)^{2k-N} \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \frac{\pi_0}{\pi_1} = \eta$$

Taking the natural logarithm of both sides and rearranging, we have

$$k \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \frac{N}{2} + \frac{1}{2} \frac{\ln(\eta)}{\ln\left(\frac{1-p}{p}\right)} = \gamma$$

In the case that both hypotheses are equally likely, the minimum probability of error decision is the "majority-vote" rule: Declare $\mathscr{H}_1$ if there are more 1s than 0s, declare $\mathscr{H}_0$ otherwise. In the event $k = \gamma$, we may decide arbitrarily; the probability of error is the same either way. Let's adopt the convention that $\mathscr{H}_0$ is declared in this case.

To compute the probability of error of the optimal rule, write

Equation:

$$\begin{aligned} P_e &= \pi_0 \Pr[\text{declare} \mathscr{H}_1 \mid \mathscr{H}_0 \text{true}] + \pi_1 \Pr[\text{declare} \mathscr{H}_0 \mid \mathscr{H}_1 \text{tr} \\ &= \pi_0 \Pr[k > \gamma \mid \mathscr{H}_0 \text{true}] + \pi_1 \Pr[k \leq \gamma \mid \mathscr{H}_1 \text{true}] \end{aligned}$$

Now k is a binomial random variable, $k \sim \text{Binomial}\,(N, \theta)$, where θ depends on which hypothesis is true. We have

Equation:

$$\begin{aligned} \Pr[k > \gamma \mid \mathscr{H}_0] &= \textstyle\sum_{k=\lfloor\gamma\rfloor+1}^{N} f_0(k) \\ &= \textstyle\sum_{k=\lfloor\gamma\rfloor+1}^{N} \binom{N}{k} p^k (1-p)^{N-k} \end{aligned}$$

and

$$\Pr[k \leq \gamma \mid \mathscr{H}_1] = \sum_{k=0}^{\lfloor\gamma\rfloor} \binom{N}{k} (1-p)^k p^{N-k}$$

Using these formulae, we may compute P_e explicitly for given values of N, p, π_0 and π_1.

MAP Interpretation

The likelihood ratio test is one way of expressing the minimum probability of error decision rule. Another way is

Rule

Declare hypothesis i such that $\pi_i f_i(\boldsymbol{x})$ is maximal.

This rule is referred to as the **maximum a posteriori**, or **MAP** rule, because the quantity $\pi_i f_i(\boldsymbol{x})$ is proportional to the posterior probability of hypothesis i. This becomes clear when we write $\pi_i = \Pr[\mathscr{H}_i]$ and $f_i(\boldsymbol{x}) = f(\boldsymbol{x}|\mathscr{H}_i)$. Then, by Bayes rule, the posterior probability of $\mathscr{H}_i$ given the data is

$$\Pr[\mathscr{H}_i \mid \boldsymbol{x}] = \frac{\Pr[\mathscr{H}_i] f(\boldsymbol{x}|\mathscr{H}_i)}{f(\boldsymbol{x})}$$

Here $f(\boldsymbol{x})$ is the unconditional density or mass function for $\boldsymbol{x}$, which is effectively a constant when trying to maximiaze with respect to i.

According to the MAP interpretation, the optimal decision boundary is the locus of points where the weighted densities (in the continuous case) $\pi_i f_i(x)$ intersect one another. This idea is illustrated in [link].

Multiple Hypotheses

One advantage the MAP formulation of the minimum probability of error decision rule has over the LRT is that it generalizes easily to M-ary hypothesis testing. If we are to choose between hypotheses $\mathscr{H}_i$, $i = \{1, \ldots, M\}$, the optimal rule is still the MAP rule

Special Case of Bayes Risk

The Bayes risk criterion for constructing decision rules assigns a cost C_{ij} to the outcome of declaring $\mathscr{H}_i$ when $\mathscr{H}_j$ is in effect. The probability of error is simply a special case of the Bayes risk corresponding to $C_{00} = C_{11} = 0$ and $C_{01} = C_{10} = 1$. Therefore, the form of the minimum probability of error decision rule is a specialization of the minimum Bayes risk decision rule: both are likelihood ratio tests. The different costs in the Bayes risk formulation simply shift the threshold to favor one hypothesis over the other.

Problems

Exercise:

Problem:

Generally speaking, when is the probability of error **zero** for the optimal rule? Phrase your answer in terms of the distributions underlying each hypothesis. Does the LRT agree with your answer in this case?

Exercise:

Problem:

Suppose we measure N independent values $x_1, \ldots, x_N$. We know the variance of our measurements ($\sigma^2 = 1$), but are unsure whether the data obeys a Laplacian or Gaussian probability law:

$$\mathscr{H}_0 : f_0(x) = \frac{1}{\sqrt{2}} e^{-\left(\sqrt{2}|r|\right)}$$

$$\mathscr{H}_1 : f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{r^2}{2}}$$

Show that the two densities have the same mean and variance, and plot the densities on the same graph.

Find the likelihood ratio.

Determine the decision regions for different values of the threshold η. Consider all possible values of $\eta > 0$

Note: There are three distinct cases.

Draw the decision regions and decision boundaries for $\eta = \left\{\frac{1}{2}, 1, 2\right\}$.

Assuming the two hypotheses are equally likely, compute the probability of error. Your answer should be a number.

Exercise:

Problem:

Arbitrary Means and Covariances

Consider the hypothesis testing problem

$$\mathscr{H}_0 : \boldsymbol{x} \sim \mathscr{N}(\mu_0, \Sigma_0)$$

$$\mathscr{H}_1 : \boldsymbol{x} \sim \mathscr{N}(\mu_1, \Sigma_1)$$

where $\mu_0 \in \mathbb{R}^d$ and $\mu_0 \in \mathbb{R}^d$, and Σ_0, Σ_1 are positive definite, symmetric d×d matrices. Write down the likelihood ratio test, and simplify, for the following cases. In each case, provide a geometric description of the decision boundary.

$\Sigma_0 = \Sigma_1$, but $\mu_0 \neq \mu_1$.

$\mu_0 = \mu_1$, but $\Sigma_0 \neq \Sigma_1$.

$\mu_0 \neq \mu_1$ and $\Sigma_0 \neq \Sigma_1$.

Exercise:

Problem:

Suppose we observe N independent realizations of a Poisson random variable k with intensity parameter λ:

$$f(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

We must decide which of two intensities is in effect:

$$\mathscr{H}_0 : \lambda = \lambda_0$$

$$\mathscr{H}_1 : \lambda = \lambda_1$$

where $\lambda_0 < \lambda_1$.

Give the minimum probability of error decision rule.

Simplify the LRT to a test statistic involving only a sufficient statistic. Apply a monotonically increasing transformation to simplify further.

Determine the distribution of the sufficient statistic under both hypotheses.

Note: Use the characteristic function to show that a sum of IID Poisson variates is again Poisson distributed.

Derive an expression for the probability of error.

Assuming the two hypotheses are equally likely, and $\lambda_0 = 5$ and $\lambda_1 = 6$, what is the minimum number N of observations needed to attain a probability of error no greater than 0.01?

Note: If you have numerical trouble, try rewriting the log-factorial so as to avoid evaluating the factorial of large integers.

Exercise:

Problem:

In [link], suppose $\pi_0 = \pi_1 = \frac{1}{2}$, and $p = 0.1$. What is the smallest value of N needed to ensure $P_e \leq 0.01$?

The Neyman-Pearson Criterion

In hypothesis testing, as in all other areas of statistical inference, there are two major schools of thought on designing good tests: Bayesian and frequentist (or classical). Consider the simple binary hypothesis testing problem

$$\mathscr{H}_0 : \boldsymbol{x} \sim f_0(\boldsymbol{x})$$

$$\mathscr{H}_1 : \boldsymbol{x} \sim f_1(\boldsymbol{x})$$

In the Bayesian setup, the prior probability $\pi_i = \Pr[\mathscr{H}_i]$ of each hypothesis occurring is assumed known. This approach to hypothesis testing is represented by the minimum Bayes risk criterion and the minimum probability of error criterion.

In some applications, however, it may not be reasonable to assign an a priori probability to a hypothesis. For example, what is the a priori probability of a supernova occurring in any particular region of the sky? What is the prior probability of being attacked by a ballistic missile? In such cases we need a decision rule that does not depend on making assumptions about the a priori probability of each hypothesis. Here the Neyman-Pearson criterion offers an alternative to the Bayesian framework.

The Neyman-Pearson criterion is stated in terms of certain probabilities associated with a particular hypothesis test. The relevant quantities are summarized in [link]. Depending on the setting, different terminology is used.

Statistics		**Signal Processing**		
Probability	**Name**	**Notation**	**Name**	**Notation**
$P_0(\text{declare}\mathscr{H}_1)$	size	α	false-alarm probability	P_F

Statistics		**Signal Processing**		
Probability	**Name**	**Notation**	**Name**	**Notation**
$P_1(\text{declare}\mathscr{H}_1)$	power	β	detection probability	P_D

Here $P_i(\text{declare}\mathscr{H}_j)$ dentoes the probability that we declare hypothesis $\mathscr{H}_j$ to be in effect when $\mathscr{H}_i$ is actually in effect. The probabilities $P_0(\text{declare}\mathscr{H}_0)$ and $P_1(\text{declare}\mathscr{H}_0)$ (sometimes called the **miss** probability), are equal to $1 - P_F$ and $1 - P_D$, respectively. Thus, P_F and P_D represent the two degrees of freedom in a binary hypothesis test. Note that P_F and P_D do not involve a priori probabilities of the hypotheses.

These two probabilities are related to each other through the decision regions. If R_1 is the decision region for $\mathscr{H}_1$, we have

$$P_F = \int f_0(\boldsymbol{x}) \text{ d } \boldsymbol{x}$$

$$P_D = \int f_1(\boldsymbol{x}) \text{ d } \boldsymbol{x}$$

The densities $f_i(\boldsymbol{x})$ are nonnegative, so as R_1 shrinks, both probabilities tend to zero. As R_1 expands, both tend to one. The ideal case, where $P_D = 1$ and $P_F = 0$, cannot occur unless the distributions do not overlap (i.e., $\int f_0(\boldsymbol{x}) f_1(\boldsymbol{x}) \text{ d } \boldsymbol{x} = 0$). Therefore, in order to increase P_D, we must also increase P_F. This represents the fundamental tradeoff in hypothesis testing and detection theory.

Example:
Consider the simple binary hypothesis test of a scalar measurement x:

$$\mathscr{H}_0 : x \sim \mathscr{N}(0, 1)$$

$$\mathscr{H}_1 : x \sim \mathscr{N}(1, 1)$$

Suppose we use a threshold test

$$x \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \gamma$$

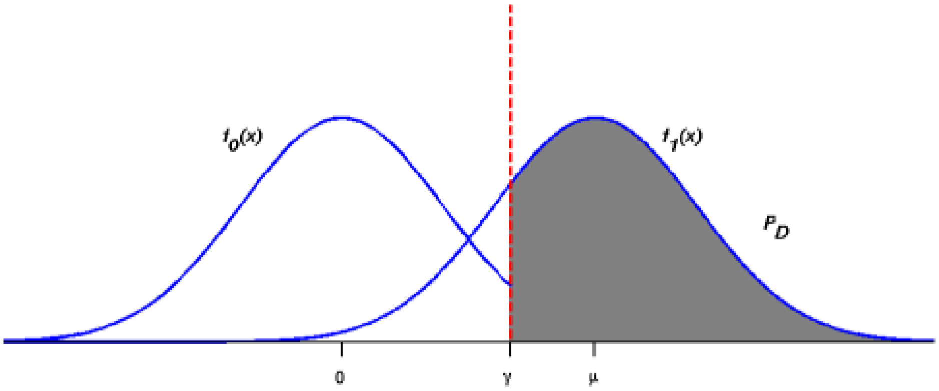
where $\gamma \in \mathbb{R}$ is a free parameter. Then the false alarm and detection probabilities are

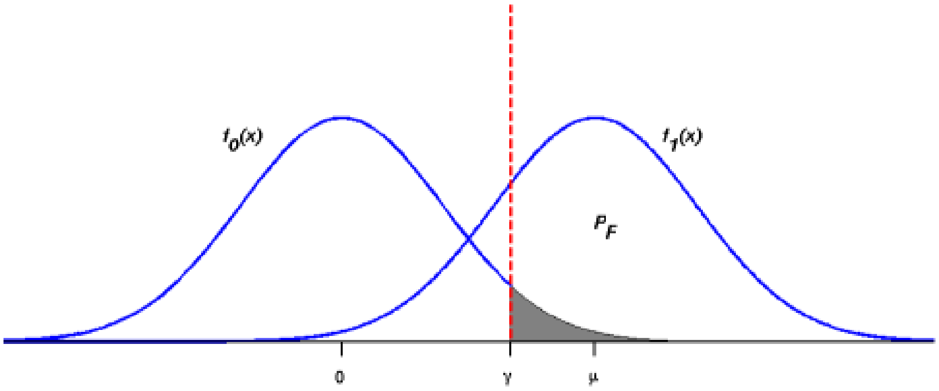
$$P_F = \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, \mathrm{d}\, t = Q(\gamma)$$

$$P_D = \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(t-1)^2}{2}} \, \mathrm{d}\, t = Q(\gamma - 1)$$

where Q denotes the Q-function. These quantities are depicted in [link].

False alarm and detection values for a certain threshold.

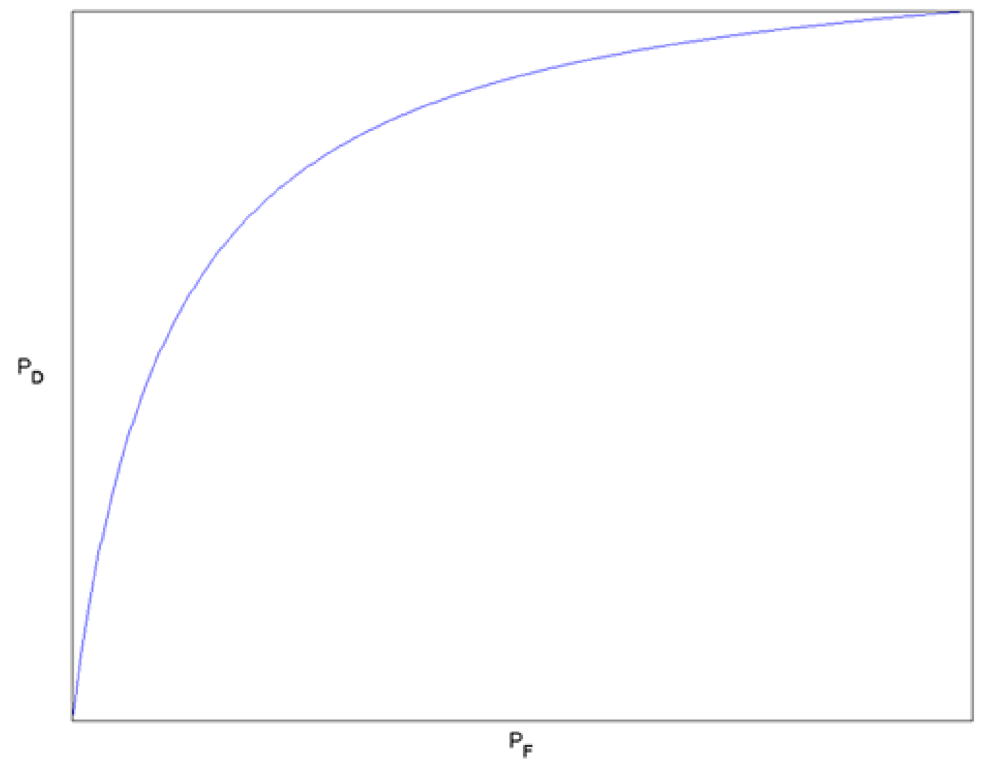




Since the Q-function is monotonicaly decreasing, it is evident that both P_D and P_F decay to zero as γ increases. There is also an explicit relationship

$$P_D = Q\left(Q^{-1}(P_F) - 1\right)$$

A common means of displaying this relationship is with a **receiver operating characteristic** (ROC) curve, which is nothing more than a plot of P_D versus P_F ([link]).



ROC curve for this example.

The Neyman-Pearson Lemma: A First Look

The Neyman-Pearson criterion says that we should construct our decision rule to have maximum probability of detection while not allowing the probability of false alarm to exceed a certain value α. In other words, the optimal detector according to the Neyman-Pearson criterion is the solution to the following constrainted optimization problem:

Neyman-Pearson Criterion

Equation:

$$\max\ \{P_D\}, \text{such that} P_F \leq \alpha$$

The maximization is over all decision rules (equivalently, over all decision regions R_0, R_1). Using different terminology, the Neyman-Pearson criterion selects the **most powerful test of size (not exceeding)** α.

Fortunately, the above optimization problem has an explicit solution. This is given by the celebrated **Neyman-Pearson lemma**, which we now state. To ease the exposition, our initial statement of this result only applies to continuous random variables, and places a technical condition on the densities. A more general statement is given later in the module.

Neyman-Pearson Lemma: initial statement

Consider the test

$$\mathscr{H}_0 : \boldsymbol{x} \sim f_0(\boldsymbol{x})$$

$$\mathscr{H}_1 : \boldsymbol{x} \sim f_1(\boldsymbol{x})$$

where $f_i(\boldsymbol{x})$ is a density. Define $\Lambda(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})}$, and assume that $\Lambda(\boldsymbol{x})$ satisfies the condition that for each $\gamma \in \mathbb{R}$, $\Lambda(\boldsymbol{x})$ takes on the value γ with probability zero under hypothesis $\mathscr{H}_0$. The solution to the optimization problem in [link] is given by

$$\Lambda(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \eta$$

where η is such that

$$P_F = \int f_0(\boldsymbol{x})\ \mathrm{d}\ \boldsymbol{x} = \alpha$$

If $\alpha = 0$, then $\eta = \infty$. The optimal test is unique up to a set of probability zero under $\mathscr{H}_0$ and $\mathscr{H}_1$.

The optimal decision rule is called the **likelihood ratio test**. $\Lambda(\boldsymbol{x})$ is the **likelihood ratio**, and η is a **threshold**. Observe that neither the likelihood ratio nor the threshold depends on the a priori probabilities $\Pr[\mathscr{H}_i]$. they depend only on the conditional densities f_i and the size constraint α. The threshold can often be solved for as a function of α, as the next example shows.

Example:
Continuing with [link], suppose we wish to design a Neyman-Pearson decision rule with size constraint α. We have
Equation:

$$\begin{aligned} \Lambda(x) &= \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-1)^2}{2}}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}} \\ &= e^{x-\frac{1}{2}} \end{aligned}$$

By taking the natural logarithm of both sides of the LRT and rarranging terms, the decision rule is not changed, and we obtain

$$x \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \ln(\eta) + \frac{1}{2} \equiv \gamma$$

Thus, the optimal rule is in fact a thresholding rule like we considered in [link]. The false-alarm probability was seen to be

$$P_F = Q(\gamma)$$

Thus, we may express the value of γ required by the Neyman-Pearson lemma in terms of α:

$$\gamma = Q^{-1}(\alpha)$$

Sufficient Statistics and Monotonic Transformations

For hypothesis testing involving multiple or vector-valued data, direct evaluation of the size (P_F) and power (P_D) of a Neyman-Pearson decision rule would require integration over multi-dimensional, and potentially complicated decision regions. In many cases, however, this can be avoided by simplifying the LRT to a test of the form

$$\boldsymbol{t} \overset{\mathscr{H}_1}{\underset{\mathscr{H}_0}{\gtrless}} \gamma$$

where the test statistic $\boldsymbol{t} = T(\boldsymbol{x})$ is a sufficient statistic for the data. Such a simplified form is arrived at by modifying both sides of the LRT with montonically increasing transformations, and by algebraic simplifications. Since the modifications do not change the decision rule, we may calculate P_F and P_D in terms of the sufficient statistic. For example, the false-alarm probability may be written

Equation:

$$\begin{aligned} P_F &= \Pr[\text{declare} \mathscr{H}_1] \\ &= \int f_0(\boldsymbol{t})\ \mathrm{d}\ \boldsymbol{t} \end{aligned}$$

where $f_0(\boldsymbol{t})$ denotes the density of $\boldsymbol{t}$ under $\mathscr{H}_0$. Since $\boldsymbol{t}$ is typically of lower dimension than $\boldsymbol{x}$, evaluation of P_F and P_D can be greatly simplified. The key is being able to reduce the LRT to a threshold test involving a sufficient statistic **for which we know the distribution**.

Example:

Common Variances, Uncommon Means

Let's design a Neyman-Pearson decision rule of size α for the problem

$$\mathscr{H}_0 : \boldsymbol{x} \sim \mathscr{N}\left(\mathbf{0}, \sigma^2 I\right)$$

$$\mathscr{H}_1 : \boldsymbol{x} \sim \mathscr{N}\left(\mu\mathbf{1}, \sigma^2 I\right)$$

where $\mu > 0$, $\sigma^2 > 0$ are known, $\mathbf{0} = (0 \ldots 0)^T$, $\mathbf{1} = (1 \ldots 1)^T$ are N-dimensional vectors, and I is the $N{\times}N$ identity matrix. The likelihood ratio is

Equation:

$$\begin{aligned} \mathit{\Lambda}(\boldsymbol{x}) &= \frac{\prod_{\mathrm{n}=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}}{\prod_{\mathrm{n}=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{{x_n}^2}{2\sigma^2}}} \\ &= \frac{e^{-\sum_{\mathrm{n}=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}}}{e^{-\sum_{\mathrm{n}=1}^{N} \frac{{x_n}^2}{2\sigma^2}}} \\ &= e^{\frac{1}{2\sigma^2} \sum_{\mathrm{n}=1}^{N} 2x_n\mu-\mu^2} \\ &= e^{\frac{1}{\sigma^2}\left(-\frac{N\mu^2}{2}+\mu \sum_{\mathrm{n}=1}^{N} x_n\right)} \end{aligned}$$

To simplify the test further we may apply the natural logarithm and rearrange terms to obtain

$$t \equiv \sum_{\mathrm{n}=1}^{N} x_n \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \frac{\sigma^2}{\mu} \ln(\eta) + \frac{N\mu}{2} \equiv \gamma$$

Note: We have used the assumption $\mu > 0$. If $\mu < 0$, then division by μ is not a monotonically increasing operation, and the inequalities would be reversed.

The test statistic t is sufficient for the unknown mean. To set the threshold γ, we write the false-alarm probability (size) as

$$P_F = \Pr[t > \gamma] = \int f_0(t) \, \mathrm{d}\, t$$

To evaluate P_F, we need to know the density of t under $\mathscr{H}_0$. Fortunately, t is the sum of normal variates, so it is again normally distributed. In particular, we have $t = \boldsymbol{Ax}$, where $\boldsymbol{A} = \mathbf{1}^T$, so

$$t \sim \mathscr{N}\left(\boldsymbol{A}\mathbf{0}, \boldsymbol{A}\left(\sigma^2 I\right)\boldsymbol{A}^T\right) = \mathscr{N}\left(0, N\sigma^2\right)$$

under $\mathscr{H}_0$. Therefore, we may write P_F in terms of the Q-function as

$$P_F = Q\left(\frac{\gamma}{\sqrt{N}\sigma}\right)$$

The threshold is thus determined by

$$\gamma = \sqrt{N}\sigma Q^{-1}(\alpha)$$

Under $\mathscr{H}_1$, we have

$$t \sim \mathscr{N}\left(\boldsymbol{A}\mathbf{1}, \boldsymbol{A}\left(\sigma^2 I\right)\boldsymbol{A}^T\right) = \mathscr{N}\left(N\mu, N\sigma^2\right)$$

and so the detection probability (power) is

$$P_D = \Pr[t > \gamma] = Q\left(\frac{\gamma - N\mu}{\sqrt{N}\sigma}\right)$$

Writing P_D as a function of P_F, the ROC curve is given by

$$P_D = Q\left(Q^{-1}(P_F) - \frac{\sqrt{N}\mu}{\sigma}\right)$$

The quantity $\frac{\sqrt{N}\mu}{\sigma}$ is called the **signal-to-noise ratio**. As its name suggests, a larger SNR corresponds to improved performance of the Neyman-Pearson decision rule.

Note: In the context of signal processing, the foregoing problem may be viewed as the problem of detecting a constant (DC) signal in additive white Gaussian noise:

$$\mathscr{H}_0 : x_n = w_n, n = 1, \ldots, N$$

$$\mathscr{H}_1 : x_n = A + w_n, n = 1, \ldots, N$$

where A is a known, fixed amplitude, and $w_n \sim \mathscr{N}\left(0, \sigma^2\right)$. Here A corresponds to the mean μ in the example.

The Neyman-Pearson Lemma: General Case

In our initial statement of the Neyman-Pearson Lemma, we assumed that for all η, the set $\{\boldsymbol{x}, \boldsymbol{x} | \, \Lambda(\boldsymbol{x}) = \eta\}$ had probability zero under $\mathscr{H}_0$. This eliminated many important problems from consideration, including tests of discrete data. In this section we remove this restriction.

It is helpful to introduce a more general way of writing decision rules. Let φ be a function of the data $\boldsymbol{x}$ with $\varphi(\boldsymbol{x}) \in [0, 1]$. φ defines the decision rule "declare $\mathscr{H}_1$ with probability $\varphi(\boldsymbol{x})$." In other words, upon observing $\boldsymbol{x}$, we flip a "$\varphi(\boldsymbol{x})$ coin." If it turns up heads, we declare $\mathscr{H}_1$; otherwise we declare $\mathscr{H}_0$. Thus far, we have only considered rules with $\varphi(\boldsymbol{x}) \in \{0, 1\}$

Neyman-Pearson Lemma

Consider the hypothesis testing problem

$$\mathscr{H}_0 : \boldsymbol{x} \sim f_0(\boldsymbol{x})$$

$$\mathscr{H}_1 : \boldsymbol{x} \sim f_1(\boldsymbol{x})$$

where f_0 and f_1 are both pdfs or both pmfs. Let $\alpha \in [0, 1)$ be the size (false-alarm probability) constraint. The decision rule

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \; \Lambda(\boldsymbol{x}) > \eta \\ \rho & \text{if } \; \Lambda(\boldsymbol{x}) = \eta \\ 0 & \text{if } \; \Lambda(\boldsymbol{x}) < \eta \end{cases}$$

is the most powerful test of size α, where η and ρ are uniquely determined by requiring $P_F = \alpha$. If $\alpha = 0$, we take $\eta = \infty$, $\rho = 0$. This test is unique up to sets of probability zero under $\mathscr{H}_0$ and $\mathscr{H}_1$.
When $\Pr[\mathit{\Lambda}(\boldsymbol{x}) = \eta] > 0$ for certain η, we choose η and ρ as follows: Write

$$P_F = \Pr[\mathit{\Lambda}(\boldsymbol{x}) > \eta] + \rho \Pr[\mathit{\Lambda}(\boldsymbol{x}) = \eta]$$

Choose η such that

$$\Pr[\mathit{\Lambda}(\boldsymbol{x}) > \eta] \leq \alpha \leq \Pr[\mathit{\Lambda}(\boldsymbol{x}) \geq \eta]$$

Then choose ρ such that

$$\rho \Pr[\mathit{\Lambda}(\boldsymbol{x}) = \eta] = \alpha - \Pr[\mathit{\Lambda}(\boldsymbol{x}) < \eta]$$

Example:

Repetition Code

Suppose we have a friend who is trying to transmit a bit (0 or 1) to us over a noisy channel. The channel causes an error in the transmission (that is, the bit is flipped) with probability p, where $0 \leq p < \frac{1}{2}$, and p is known. In order to increase the chance of a successful transmission, our friend sends the same bit N times. Assume the N transmissions are statistically independent. Under these assumptions, the bits you receive are Bernoulli random variables: $x_n \sim \text{Bernoulli}\,(\theta)$. We are faced with the following hypothesis test:

$$\mathscr{H}_0 : \theta = p (0 \text{ sent})$$

$$\mathscr{H}_1 : \theta = 1 - p (1 \text{ sent})$$

We decide to decode the received sequence $\boldsymbol{x} = (x_1 \ldots x_N)^T$ by designing a Neyman-Pearson rule. The likelihood ratio is
Equation:

$$\begin{aligned}\Lambda(\boldsymbol{x}) &= \frac{\prod_{n=1}^{N}(1-p)^{x_n}p^{1-x_n}}{\prod_{n=1}^{N}p^{x_n}(1-p)^{1-x_n}}\\ &= \frac{(1-p)^{k}p^{N-k}}{p^{k}(1-p)^{N-k}}\\ &= \left(\frac{1-p}{p}\right)^{2k-N}\end{aligned}$$

where $k = \sum_{n=1}^{N} x_n$ is the number of 1s received.

Note:k is a sufficient statistic for θ.

The LRT is

$$\left(\frac{1-p}{p}\right)^{2k-N} \overset{\mathscr{H}_1}{\underset{\mathscr{H}_0}{\gtrless}} \eta$$

Taking the natural logarithm of both sides and rearranging, we have

$$k \overset{\mathscr{H}_1}{\underset{\mathscr{H}_0}{\gtrless}} \frac{N}{2} + \frac{1}{2}\frac{\ln(\eta)}{\ln\left(\frac{1-p}{p}\right)} = \gamma$$

The false alarm probability is
Equation:

$$\begin{aligned}P_F &= \Pr[k > \gamma] + \rho\Pr[k = \gamma]\\ &= \sum_{k=\gamma+1}^{N}\binom{N}{k}p^{k}(1-p)^{N-k} + \rho\binom{N}{\gamma}p^{\gamma}(1-p)^{N-\gamma}\end{aligned}$$

γ and ρ are chosen so that $P_F = \alpha$, as described above.

The corresponding detection probability is

Equation:

$$\begin{aligned} P_D &= \Pr[k > \gamma] + \rho \Pr[k = \gamma] \\ &= \sum_{k=\gamma+1}^{N} \binom{N}{k} (1-p)^k p^{N-k} + \rho \binom{N}{\gamma} (1-p)^\gamma p^{N-\gamma} \end{aligned}$$

Problems

Exercise:

Problem:

Design a hypothesis testing problem involving continous random variables such that $\Pr[\Lambda(x) = \eta] > 0$ for certain values of η. Write down the false-alarm probability as a function of the threshold. Make as general a statement as possible about when the technical condition is satisfied.

Exercise:

Problem:

Consider the scalar hypothesis testing problem

$$\mathscr{H}_0 : x \sim f_0(x)$$

$$\mathscr{H}_1 : x \sim f_1(x)$$

where

$$f_i(x) = \frac{1}{\pi\left(1 + (x-i)^2\right)}, i = \{0, 1\}$$

Write down the likelihood ratio test.

Determine the decision regions as a function of η_1 for all $\eta > 0$. Draw a representative of each. What are the "critical" values of η?

Note: There are five distinct cases.

Compute the size and power (P_F and P_D) in terms of the threshold η_1 and plot the ROC.

Note:

$$\int \frac{1}{1+x^2} \mathrm{d}\, x = \arctan(x)$$

Suppose we decide to use a simple threshold test $x \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta$ instead of the Neyman-Pearson rule. Does our performance $\mathcal{H}_0$ suffer much? Plot the ROC for this decision rule on the same graph as for the previous ROC.

Exercise:

Problem:

Suppose we observe N independent realizations of a Poisson random variable k with intensity parameter λ:

$$f(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

We must decide which of two intensities is in effect:

$$\mathcal{H}_0 : \lambda = \lambda_0$$

$$\mathcal{H}_1 : \lambda = \lambda_1$$

where $\lambda_0 < \lambda_1$.

Write down the likelihood ratio test.

Simplify the LRT to a test statistic involving only a sufficient statistic. Apply a monotonically increasing transformation to simplify further.

Determine the distribution of the sufficient statistic under both hypotheses.

Note: Use the characteristic function to show that a sum of IID Poisson variates is again Poisson distributed.

Derive an expression for the probability of error.

Assuming the two hypotheses are equally likely, and $\lambda_0 = 5$ and $\lambda_1 = 6$, what is the minimum number N of observations needed to attain a false-alarm probability no greater than 0.01?

Note: If you have numerical trouble, try rewriting the log-factorial so as to avoid evaluating the factorial of large integers.

Exercise:

Problem:

In [link], suppose $p = 0.1$. What is the smallest value of N needed to ensure $P_F \leq 0.01$? What is P_D in this case?

The Minimum Variance Unbiased Estimator

In Search of a Useful Criterion

In parameter estimation, we observe an N-dimensional vector $\boldsymbol{X}$ of measurements. The distribution of $\boldsymbol{X}$ is governed by a density or probability mass function f$_\theta$ $(\boldsymbol{x})$, which is parameterized by an unknown parameter $\boldsymbol{\theta}$. We would like to establish a useful criterion for guiding the design and assessing the quality of an estimator $\widehat{\theta(\boldsymbol{x})}$. We will adopt a classical (frequentist) view of the unknown parameter: it is not itself random, it is simply unknown.

One possibility is to try to design an estimator that minimizes the **mean-squared error**, that is, the expected squared deviation of the estimated parameter value from the true parameter value. For a scalar parameter, the MSE is defined by
Equation:

$$\mathrm{MSE}\left(\hat{\theta},\theta\right)=E\left[\left(\widehat{\theta(\boldsymbol{x})}-\theta\right)^{2}\right]$$

For a vector parameter $\boldsymbol{\theta}$, this definition is generalized by
Equation:

$$\mathrm{MSE}\left(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}\right)=E\left[\left(\widehat{\theta(\boldsymbol{x})}-\boldsymbol{\theta}\right)^{T}\left(\widehat{\theta(\boldsymbol{x})}-\boldsymbol{\theta}\right)\right]$$

The expectation is with respect to the distribution of $\boldsymbol{X}$. Note that for a given estimator, the MSE is a function of $\boldsymbol{\theta}$.

While the MSE is a perfectly reasonable way to assess the quality of an estimator, it does not lead to a useful design criterion. Indeed, the estimator that minimizes the MSE is simply the estimator
Equation:

$$\widehat{\theta(\boldsymbol{x})} = \boldsymbol{\theta}$$

Unfortunately, this depends on the value of the unknown parameter, and is therefore not realizeable! We need a criterion that leads to a realizeable estimator.

Note: In the Bayesian Approach to Parameter Estimation, the MSE **is** a useful design rule.

The Bias-Variance Decomposition of the MSE

It is possible to rewrite the MSE in such a way that a useful optimality criterion for estimation emerges. For a scalar parameter θ, [Insert 1] This expression is called the **bias-variance decomposition** of the mean-squared error. The first term on the right-hand side is called the **variance** of the estimator, and the second term on the right-hand side is the square of the **bias** of the estimator. The formal definition of these concepts for vector parameters is now given:

Let $\widehat{\boldsymbol{\theta}}$ be an estimator of the parameter θ.

variance

The **variance** of $\widehat{\boldsymbol{\theta}}$ is [Insert 2]

bias

The **bias** of $\widehat{\boldsymbol{\theta}}$ is [Insert 3]

The bias-variance decomposition also holds for vector parameters: [Insert 4] The proof is a straighforward generalization of the argument for the scalar parameter case.

Exercise:

Problem:

Prove the bias-variance decomposition of the MSE for the vector parameter case.

The Bias-Variance Tradeoff

The MSE decomposes into the sum of two non-negative terms, the squared bias and the variance. In general, for an arbitrary estimator, both of these terms will be nonzero. Furthermore, as an estimator is modified so that one term increases, typically the other term will decrease. This is the so-called **bias-variance tradeoff**. The following example illustrates this effect.

Example:
Let $\tilde{A} = \alpha \frac{1}{N} \sum_{n=1}^{N} x_n$, where $x_n = A + w_n$, $w_n \sim \mathcal{N}\left(0, \sigma^2\right)$, and α is an arbitrary constant.
Let's find the value of α that minimizes the MSE.
Equation:

$$\text{MSE}\left(\tilde{A}\right) = E\left[\left(\tilde{A} - A\right)^2\right]$$

Note: $\tilde{A} = \alpha S_N$, $S_N \sim \mathcal{N}\left(\text{A}, \frac{\sigma^2}{N}\right)$

Equation:

$$\begin{aligned} \mathrm{MSE}\left(\tilde{A}\right) &= E\left[\tilde{A}^{2}\right] - 2E\left[\tilde{A}\right]A + A^2 \\ &= \alpha^2 E\left[\tfrac{1}{N^2}\textstyle\sum_{i,j=1}^{N} x_i x_j\right] - 2\alpha E\left[\tfrac{1}{N}\textstyle\sum_{n=1}^{N} x_n\right]A + A^2 \\ &= \alpha^2 \tfrac{1}{N^2}\textstyle\sum_{i,j=1}^{N} E[x_i x_j] - 2\alpha \times \tfrac{1}{N}\textstyle\sum_{n=1}^{N} E[x_n] + A^2 \end{aligned}$$

$$E[x_i x_j] = \begin{cases} A^2 + \sigma^2 \text{ if } i = j \\ A^2 \text{ if } i \neq j \end{cases}$$

Equation:

$$\begin{aligned} \mathrm{MSE}\left(\tilde{A}\right) &= \alpha^2\left(A^2 + \tfrac{\sigma^2}{N}\right) - 2\alpha A^2 + A^2 \\ &= \tfrac{\alpha^2\sigma^2}{N} + (\alpha - 1)^2 A^2 \end{aligned}$$

$$\sigma\left(\tilde{A}\right)^2 = \frac{\alpha^2\sigma^2}{N}$$

$$\mathrm{Bias}^2\left(\tilde{A}\right) = (\alpha - 1)^2 A^2$$

$$\frac{\partial\ \mathrm{MSE}\left(\tilde{A}\right)}{\partial\alpha} = \frac{2\alpha\sigma^2}{N} + 2\,(\alpha - 1)A^2 = 0$$

Equation:

$$\alpha^* = \frac{A^2}{A^2 + \frac{\sigma^2}{N}}$$

The optimal value α^* dpends on the unknown parameter A! Therefore the estimator is not realizable.
Note that the problematic dependence on the parameter enters through the Bias component of the MSE. Therefore, a reasonable alternative is to constrain the estimator to be unbiased, and then find the estimator that

produces the minimum variance (and hence provides the minimum MSE among all unbiased estimators).

Note: Sometimes no unbiased estimator exists, and we cannot proceed at all in this direction.

In this example, note that as the value of α varies, one of the squared bias or variance terms increases, while the other one decreases. Futhermore, note that **the dependence of the MSE on the unknown parameter is manifested in the bias**.

Unbiased Estimators

Since the bias depends on the value of the unknown parameter, it seems that any estimation criterion that depends on the bias would lead to an unrealizable estimator, as the previous example suggests (although in certain cases realizable minimum MSE estimators can be found). As an alternative to minimizing the MSE, we could focus on estimators that have a bias of zero. In this case, the bias contributes zero to the MSE, and in particular, it does not involve the unknown parameter. By focusing on estimators with zero bias, we may hope to arrive at a design criterion that yields **realizable** estimators.

unbiased

> An estimator $\widehat{\boldsymbol{\theta}}$ is called **unbiased** if its bias is zero for all values of the unknown parameter. Equivalently, [Insert 5]

For an estimator to be unbiased we require that **on average** the estimator will yield the true value of the unknown parameter. We now give some examples.

The sample mean of a random sample is always an unbiased estimator for the mean.

Example:
Estimate the DC level in the **Guassian white noise**.
Suppose we have data $x_1, \ldots, x_N$ and model the data by

$$\forall n, n \in \{1, \ldots, N\} : (x_n = A + w_n)$$

where A is the unknown DC level, and $w_n \sim \mathcal{N}\left(\sigma, \sigma^2\right)$.
The parameter is $-\infty < A < \infty$.
Consider the **sample-mean estimator**:

$$\widehat{A} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

Is $\widehat{A}$ unbiased? **Yes.**
Since $E[\cdot]$ is a linear operator,

$$E\left[\widehat{A}\right] = \frac{1}{N} \sum_{n=1}^{N} E[x_n] = \frac{1}{N} \sum_{n=1}^{N} A = A$$

Therefore, A is unbiased!
What does the unbiased restriction really imply? Recall that $\hat{\theta} = g(x)$, a function of the data. Therefore,

$$\forall \theta : \left(E\left[\hat{\theta}\right] = \theta\right)$$

and

$$\forall \theta : \left(E\left[\hat{\theta}\right] = \int g(x)\ \mathrm{p}\ (x\,|\,\theta)\ \mathrm{d}\ x = \theta\right)$$

Hence, to be unbiased, the estimator ($g(\cdot)$) must satisfy an integral equation involving the densities p $(x\,|\,\theta)$.
It is possible that an estimator can be unbiased for some parameter values, but be biased for others.

The bias of an estimator may be zero for **some** values of the unknown parameter, but not others. In this case, the estimator is **not** an unbiased estimator.

Example:

$$\tilde{A} = \frac{1}{2N}\sum_{n=1}^{N} x_n$$

$$E\left[\tilde{A}\right] = \frac{1}{2}A = \begin{cases} 0 \text{ if } & (A = 0) \Rightarrow \text{unbiased} \\ \frac{1}{2}A \text{ if } & (A \neq 0) \Rightarrow \text{biased} \end{cases}$$

An unbiased estimator is not necessarily a good estimator.

Some unbiased estimators are more useful than others.

Example:

$$\forall w_n, w_n \sim \mathscr{N}\left(\sigma, \sigma^2\right) : (x_n = A + w_n)$$

$$\widehat{A}_1 = x_1$$

$$E\left[\widehat{A}_1\right] = A$$

$$\widehat{A}_2 = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$E\left[\widehat{A}_2\right] = A$$

$$\sigma\left(\widehat{A}_1\right)^2 = \sigma^2$$

$$\sigma\left(\widehat{A}_2\right)^2 = \frac{\sigma^2}{N}$$

Both estimators are unbiased, but $\widehat{A}_2$ has a much lower variance and therefore is a better estimator.

Note: $\widehat{A_1(N)}$ is inconsistent. $\widehat{A_2(N)}$ is consistent.

Minimum Variance Unbiased Estimators

Direct minimization of the MSE generally leads to non-realizable estimators. Since the dependence of an estimator on the unknown parameter appears to come from the bias term, we hope that constraining the bias to be zero will lead to a useful design criterion. But if the bias is zero, then the mean-squared error is just the variance. This gives rise to the **minimum variance unbiased estimator (MVUE)** for $\boldsymbol{\theta}$.

MVUE

> An estimator $\widehat{\boldsymbol{\theta}}$ is the **minimum variance unbiased estimator** if it is unbiased and has the smallest variance of any unbiased estimator for

all values of the unknown parameter. In other words, the MVUE satisfies the following two properties: [Insert 6]

The minimum variance unbiased criterion is the primary estimation criterion in the classical (non-Bayesian) approach to parameter estimation. Before delving into ways of finding the MVUE, let's first consider whether the MVUE always exists.

Existence of the MVUE

The MVUE does not always exist. In fact, it may be that no unbiased estimators exist, as the following example demonstrates.

Place [Insert 7] here and make it an example (5).

Even if unbiased estimators exist, it may be that no single unbiased estimator has the minimum variance for **all** values of the unknown parameter.

Place [Insert 8] here and make it an example (6).

Exercise:

Problem:

Compute the variances of the estimators in the previous examples. Using the Cramer-Rao Lower bound, show that one of these two estimators has minimum variance among all unbiased estimators. Deduce that no single realizable estimator can have minimum variance among all unbiased estimators for all parameter values (i.e., the MVUE does not exist). When using the Cramer-Rao bound, note that the likelihood is not differentable at $\theta = 0$.

Methods for Finding the MVUE

Despite the fact that the MVUE doesn't always exist, in many cases of interest it does exist, and we need methods for finding it. Unfortunately, there is no 'turn the crank' algorithm for finding MVUE's. There are, instead, a variety of techniques that can **sometimes** be applied to find the MVUE. These methods include:

1. Compute the Cramer-Rao Lower Bound, and check the condition for equality.
2. Find a **complete** sufficient statistic and apply the Rao-Blackwell Theorem.
3. If the data obeys a general linear model, restrict to the class of linear unbiased estimators, and find the minimum variance estimator within that class. This method is in general suboptimal, although when the noise is Gaussian, it produces the MVUE.

The Cramer-Rao Lower Bound

The **Cramer-Rao Lower Bound** (CRLB) sets a lower bound on the variance of **any** unbiased estimator. This can be extremely useful in several ways:

1. If we find an estimator that achieves the CRLB, then we know that we have found an MVUB estimator!
2. The CRLB can provide a benchmark against which we can compare the performance of any unbiased estimator. (We know we're doing very well if our estimator is "close" to the CRLB.)
3. The CRLB enables us to rule-out impossible estimators. That is, we know that it is physically impossible to find an unbiased estimator that beats the CRLB. This is useful in feasibility studies.
4. The theory behind the CRLB can tell us if an estimator exists that achieves the bound.

Estimator Accuracy

Consider the likelihood function $\mathrm{p}\left(x \mid \theta\right)$, where θ is a scalar unknown (parameter). We can plot the likelihood as a function of the unknown, as shown in [link].

[missing_resource:]

The more "peaky" or "spiky" the likelihood function, the easier it is to determind the unknown parameter.

Example:
Suppose we observe

$$x = A + w$$

where $w \sim \mathcal{N}\left(\sigma, \sigma^2\right)$ and A is an unknown parameter. The "smaller" the noise w is, the easier it will be to estimate A from the observation x.
Suppose $A = 3$ and $\sigma = 1/3$.
[missing_resource:]
Given this density function, we can easily rule-out estimates of A greater than 4 or less than 2, since it is very unlikely that such A could give rise to out observation.
On the other hand, suppose $\sigma = 1$.
[missing_resource:]
In this case, it is very difficult to estimate A. Since the noise power is larger, it is very difficult to distinguish A from the noise.
The key thing to notice is that the estimation accuracy of A depends on σ, which in effect determines the peakiness of the likelihood. The more peaky, the better localized the data is about the true parameter.
To quantify the notion, note that the peakiness is effectively measured by the negative of the second derivative of the log-likelihood at its peak, as seen in [link].
[missing_resource:]

Example:

$$x = A + w$$

Equation:

$$\log \mathrm{p}\left(x \mid A\right) = \left(-\log \sqrt{2\pi\sigma^{2}}\right) - \frac{1}{2\sigma^{2}}\left(x - A\right)^{2}$$

$$\frac{\partial \log \mathrm{p}\left(x \mid A\right)}{\partial A} = \frac{1}{\sigma^{2}}\left(x - A\right)$$

Equation:

$$-\frac{\partial^{2}\log \mathrm{p}\left(x \mid A\right)}{\partial \text{msup}} = \frac{1}{\sigma^{2}}$$

The curvature increases as σ^{2} decreases (curvature=peakiness).
In general, the curavture will depend on the observation data; $-\frac{\partial^{2}\log\mathrm{p}(x \mid A)}{\partial \text{msup}}$ is a function of x.
Therefore, an average measure of curvature is more appropriate.

Equation:

$$-E\left[\frac{\partial^{2}\log \mathrm{p}\left(x \mid \theta\right)}{\partial \text{msup}}\right]$$

This average-out randomness due to the data and is a function of θ alone.

We are now ready to state the CRLB theorem.
Cramer-Rao Lower Bound Theorem

Assume that the pdf $\mathrm{p}\left(x \mid \theta\right)$ satisfies the "regularity" condition

$$\forall \theta : \left(E\left[\frac{\partial \log \mathrm{p}\left(x \mid \theta\right)}{\partial \theta}\right] = 0\right)$$

where the expectation is take with respect to $\mathrm{p}\left(x \mid \theta\right)$. Then, the variance of any unbiased estimator $\hat{\theta}$ must satisfy

Equation:

$$\sigma\left(\hat{\theta}\right)^{2} \geq \frac{1}{-E\left[\frac{\partial^{2}\log\mathrm{p}(x \mid \theta)}{\partial \text{msup}}\right]}$$

where the derivative is evaluated at the true value of θ and the expectation is with respect to $\mathrm{p}\,(x\,|\,\theta)$. Moreover, an unbiased estimator may be found that attains the bound for all θ if and only if

Equation:

$$\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial \theta} = I(\theta)\,(g(\theta) - \theta)$$

for some functions g and I.

The corresponding estimator is MVUB and is given by $\hat{\theta} = g(x)$, and the minimum variance is $\frac{1}{I(\theta)}$.

Example:

$$x = A + w$$

where

$$w \sim \mathcal{N}\left(0, \sigma^2\right)$$

$$\theta = A$$

$$\forall A : \left(E\left[\frac{\partial \log p}{\partial \theta}\right] = E\left[\frac{1}{\sigma^2}\,(x - A)\right] = 0\right)$$

$$\mathrm{CRLB} = \frac{1}{-E\left[\frac{\partial^2 \log p}{\partial \text{msup}}\right]} = \frac{1}{\frac{1}{\sigma^2}} = \sigma^2$$

Therefore, any unbiased estimator $\widehat{A}$ has $\text{σ}\left(\widehat{A}\right)^2 \geq \sigma^2$. But we know that $\widehat{A} = x$ has $\text{σ}\left(\widehat{A}\right)^2 = \sigma^2$. Therefore, $\widehat{A} = x$ is the MVUB estimator.

Note:

$$\theta = A$$

$$I(\theta) = \frac{1}{\sigma^2}$$

$$g(x) = x$$

First consider the reguarity condition:

$$E\left[\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\right] = 0$$

Note:

$$E\left[\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\right] = \int \frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,\theta = \int \frac{\partial\ \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \mathrm{d}\,\theta$$

Now assuming that we can interchange order of differentiation and integration

$$E\left[\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\right] = \frac{\partial \int\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,\theta}{\partial\theta} = \frac{\partial 1}{\partial\theta} = 0$$

So the regularity condition is satisfied whenever this interchange is possible[footnote]; i.e., when derivative is well-defined, fails for uniform density.
This is simply the **Fundamental Theorem of Calculus** applied to $\mathrm{p}\,(x\,|\,\theta)$. So long as $\mathrm{p}\,(x\,|\,\theta)$ is absolutely continuous with respect to the **Lebesgue measure**, this is possible.

Now lets derive the CRLB for a scalar parameter $\alpha = g(\theta)$, where the pdf is $\mathrm{p}\,(x\,|\,\theta)$. Consider any unbiased estimator of α:

$$\widehat{\alpha} \in \left(E\left[\widehat{\alpha}\right] = \alpha = g(\theta)\right)$$

Note that this is equivalent to

$$\int \widehat{\alpha}\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,x = g(\theta)$$

where $\widehat{\alpha}$ is unbiased. Now differentiate both side

$$\int \widehat{\alpha} \frac{\partial\ \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \mathrm{d}\,x = \frac{\partial\, g(\theta)}{\partial\theta}$$

or

$$\int \widehat{\alpha} \frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,x = \frac{\partial\, g(\theta)}{\partial\theta}$$

Now, exmploiting the regularity condition,
Equation:

$$\int\ (\widehat{\alpha} - \alpha) \frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,x = \frac{\partial\, g(\theta)}{\partial\theta}$$

since

$$\int \alpha \frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,x = \alpha E[\log \mathrm{p}\,(x\,|\,\theta)] = 0$$

Now apply the **Cauchy-Schwarz inequality** to the integral above:

$$\left(\frac{\partial\, g(\theta)}{\partial\theta}\right)^2 = \left(\int\ (\widehat{\alpha}-\alpha)\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,x\right)^2$$

$$\left(\frac{\partial\, g(\theta)}{\partial\theta}\right)^2 \le \int\ (\widehat{\alpha}-\alpha)^2\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,x \int\ \left(\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\right)^{\mathrm{p}(x|\theta)}\ \ \mathrm{d}\ \theta$$

σ$(\widehat{\alpha})^2$ is $\int\ (\widehat{\alpha}-\alpha)^2\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\ x$, so

Equation:

$$\sigma(\widehat{\alpha})^2 \ge \frac{\left(\frac{\partial\, g(\theta)}{\partial\theta}\right)^2}{E\left[\left(\frac{\partial \log \mathrm{p}(x|\theta)}{\partial\theta}\right)^2\right]}$$

Now we note that

$$E\left[\left(\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 \log \mathrm{p}\,(x\,|\,\theta)}{\partial \mathrm{msup}}\right]$$

Why? Regularity condition.

$$E\left[\frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\right] = \int \frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,x = 0$$

Thus,

$$\frac{\partial \int \frac{\partial \log \mathrm{p}(x|\theta)}{\partial\theta}\ \mathrm{p}\,(x\,|\,\theta)\ \mathrm{d}\,x}{\partial\theta} = 0$$

or

$$\int \frac{\partial^2 \log \mathrm{p}\,(x\,|\,\theta)}{\partial \mathrm{msup}}\ \mathrm{p}\,(x\,|\,\theta) + \frac{\partial \log \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \frac{\partial\ \mathrm{p}\,(x\,|\,\theta)}{\partial\theta}\ \mathrm{d}\,x = 0$$

Therefore,

$$-E\left[\frac{\partial^2 \log \mathrm{p}\left(x|\theta\right)}{\partial \text{msup}}\right] = \int \frac{\partial \log \mathrm{p}\left(x|\theta\right)}{\partial\theta}\frac{\partial \log \mathrm{p}\left(x|\theta\right)}{\partial\theta}\mathrm{p}\left(x|\theta\right)\,\mathrm{d}\,x = E\left[\left(\frac{\partial \log \mathrm{p}\left(x|\theta\right)}{\partial\theta}\right)^2\right]$$

Thus, [link] becomes

$$\mathrm{σ}\left(\widehat{\alpha}\right)^2 \geq \frac{\left(\frac{\partial g(\theta)}{\partial\theta}\right)^2}{-E\left[\frac{\partial^2 \log \mathrm{p}(x|\theta)}{\partial \text{msup}}\right]}$$

Note: If $g(\theta) = \theta$, then numerator is 1.

Example:DC Level in White Guassian Noise

$$\forall n, n \in \{1, \ldots, N\} : (x_n = A + w_n)$$

where

$$w_n \sim \mathcal{N}\left(0, \sigma^2\right)$$

$$\mathrm{p}\left(x|A\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}}e^{-\left(\frac{1}{\sigma^2}\sum_{n=1}^{N}\left(x_n - A\right)^2\right)}$$

$$\frac{\partial \log \mathrm{p}\left(x|A\right)}{\partial A} = \frac{\partial\left(\left(-\log\left(2\pi\sigma^2\right)^{\frac{N}{2}}\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(x_n - A\right)^2\right)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n - A$$

$$E\left[\frac{\partial \log \mathrm{p}\left(x|A\right)}{\partial A}\right] = 0$$

$$\frac{\partial^2 \log \mathrm{p}\left(x|A\right)}{\partial \text{msup}} = -\frac{N}{\sigma^2}$$

Therefore, the variance of any unbiased estimator satisfies:

$$\mathrm{σ}\left(\widehat{A}\right)^2 \geq \frac{\sigma^2}{N}$$

The sample-mean estimator $\widehat{A} = \frac{1}{N}\sum_{n=1}^{N} x_n$ attains this bound and therefore is MVUB.

Corollary

When the CRLB is attained

$$\sigma\left(\hat{\theta}\right)^{2} = \frac{1}{I(\theta)}$$

where

$$I(\theta) = -E\left[\frac{\partial^{2}\log\,\mathrm{p}\,(x\,|\,\theta)}{\partial\text{msup}}\right]$$

The quantity $I(\theta)$ is called **Fisher Information** that x contains about θ.

By CRLB Theorem,

$$\sigma\left(\hat{\theta}\right)^{2} = \frac{1}{-E\left[\frac{\partial^{2}\mathrm{logp}(x\,|\,\theta)}{\partial\text{msup}}\right]}$$

and

$$\frac{\partial\log\,\mathrm{p}\,(x\,|\,\theta)}{\partial\theta} = I(\theta)\left(\hat{\theta} - \theta\right)$$

This yields

$$\frac{\partial^{2}\log\,\mathrm{p}\,(x\,|\,\theta)}{\partial\text{msup}} = \frac{\partial I(\theta)}{\partial\theta}\left(\hat{\theta} - \theta\right) - I(\theta)$$

which in turn yields

$$-E\left[\frac{\partial^{2}\log\,\mathrm{p}\,(x\,|\,\theta)}{\partial\text{msup}}\right] = I(\theta)$$

So,

$$\sigma\left(\hat{\theta}\right)^{2} = \frac{1}{I(\theta)}$$

The CRLB is not always attained.

Example:
Phase Estimation

$$\forall n, n \in \{1, \ldots, N\} : (x_n = A\cos(2\pi f_0 n + \varphi) + w_n)$$

The amplitude and frequency are assumed known

$$w_n \sim \mathcal{N}\left(0, \sigma^{2}\right)$$

idd.

$$\mathrm{p}\left(x \,|\, \varphi\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} e^{-\left(\frac{1}{2\sigma^2}\sum_{n=1}^{N} x_n - A\cos(2\pi f_0 n + \varphi)\right)}$$

$$\frac{\partial \log \mathrm{p}\left(x \,|\, \varphi\right)}{\partial \varphi} = \left(-\frac{A}{\sigma^2}\right) \sum_{n=1}^{N} x_n \sin(2\pi f_0 n + \varphi) - \frac{A}{2} \sin(4\pi f_0 n + \varphi)$$

$$\frac{\partial^2 \log \mathrm{p}\left(x \,|\, \varphi\right)}{\partial \text{msup}} = \left(-\frac{A}{\sigma^2}\right) \sum_{n=1}^{N} x_n \cos(2\pi f_0 n + \varphi) - A\cos(2\pi f_0 n + 2\varphi)$$

$$-E\left[\frac{\partial^2 \log \mathrm{p}\left(x \,|\, \varphi\right)}{\partial \text{msup}}\right] = \frac{A^2}{\sigma^2} \sum_{n=1}^{N} 1/2 + 1/2\cos(4\pi f_0 n + 2\varphi) - \cos(4\pi f_0 n + 2\varphi)$$

Since $I(\varphi) = -E\left[\frac{\partial^2 \log \mathrm{p}(x \,|\, \varphi)}{\partial \text{msup}}\right]$,

$$I(\varphi) \simeq \frac{NA^2}{2\sigma^2}$$

because $\forall f_0, 0 < f_0 < k : \left(\frac{1}{N}\sum \cos(4\pi f_0 n) \simeq 0\right)$ Therefore,

$$\sigma\left(\widehat{\varphi}\right)^2 \geq \frac{2\sigma^2}{NA^2}$$

In this case, it can be shown that there does not exist a g such that

$$\frac{\partial \log \mathrm{p}\left(x \,|\, \varphi\right)}{\partial \varphi} \neq I(\varphi)\left(g(x) - \varphi\right)$$

Therefore, an unbiased phase estimator that attains the CRLB does not exist.
However, a MVUB estimator may still exist--only its variance will be larger than the CRLB.

Efficiency

An estimator which is unbiased and attains the CRLB is said to be **efficient**.

Example:
Sample-mean estimator is efficient.

Example:
Supposed three unbiased estimators exist for a param θ.
[missing_resource:]
[missing_resource:]

Example:
Sinusoidal Frequency Estimation

$$\forall f_0, 0 < f_0 < 1/2 : (s_n(f_0) = A\cos(2\pi f_0 n + \varphi))$$

$$\forall n, n \in \{1, \ldots, N\} : (x_n = s_n(f_0) + w_n)$$

A and φ are known, while f_0 is unknown.

$$\sigma\left(\widehat{f_0}\right)^2 \geq \frac{\sigma^2}{A^2 \sum_{n=1}^{N} \left(2\pi n \sin(2\pi f_0 n + \varphi)\right)^2}$$

Suppose $\frac{A^2}{\sigma^2} = 1$ (SNR), where $N = 10$ and $\varphi = 0$.
[missing_resource:]

Some frequencies are easier to estimator (lower CRLB, but not necessarily just lower bound) than others.

CRLB for Vector Parameter

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}$$

$\widehat{\boldsymbol{\theta}}$ is unbiased, i.e.,

$$\forall i, i \in \{1, \ldots, p\} : \left(E\left[\hat{\theta}_i\right] = \theta_i\right)$$

CRLB

$$\sigma\left(\hat{\theta}_i\right)^2 \geq \left(I(\theta)\right)^{-1}{}_{i,i}$$

where

$$\forall i \wedge j : \left(I(\theta)_{i,j} = -E\left[\frac{\partial^2 \log \mathrm{p}\ (\boldsymbol{x}\,|\,\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \right] \right)$$

$I(\theta)$ is the **Fisher Information Matrix**.

Cramer-Rao Lower Bound - Vector Parameter

Assume the pdf $\mathrm{p}\ (\boldsymbol{x}\,|\,\boldsymbol{\varphi})$ satisfies the "regularity" condition

$$\forall \boldsymbol{\theta} : \left(E\left[\frac{\partial \log \mathrm{p}\ (\boldsymbol{x}\,|\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] = 0 \right)$$

Then the convariance matrix of any unbiased estimator $\hat{\theta}$ satisfies

$$C_{\hat{\theta}} - (I(\boldsymbol{\theta}))^{-1} \geq 0$$

(meaning $C_{\hat{\theta}} - (I(\boldsymbol{\theta}))^{-1}$ is p.s.d.) The Fisher Information matrix is

$$I(\boldsymbol{\theta})_{i,j} = -E\left[\frac{\partial^2 \log \mathrm{p}\ (\boldsymbol{x}\,|\,\boldsymbol{\theta})}{\partial \text{msup}} \right]$$

Furthermore, $\hat{\theta}$ attains the CRLB ($C_{\hat{\theta}} = (I(\boldsymbol{\theta}))^{-1}$) iff

$$\frac{\partial \log \mathrm{p}\ (\boldsymbol{x}\,|\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = I(\boldsymbol{\theta})\,(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{\theta})$$

and

$$\hat{\theta} = \boldsymbol{g}(\boldsymbol{x})$$

Example:DC Level in White Guassian Noise

$$\forall n, n \in \{1, \ldots, N\} : (x_n = A + w_n)$$

A is unknown and $w_n \sim \mathcal{N}\left(0, \sigma^2\right)$, where σ^2 is unknown.

$$\boldsymbol{\theta} = \begin{pmatrix} A \\ \sigma^2 \end{pmatrix}$$

$$\log \mathrm{p}\ (\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \left(-\left(\frac{N}{2} \log\left(2\pi\right) \right) \right) - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=2}^{N} (x_n - A)^2$$

$$\frac{\partial \log \mathrm{p}\ (\boldsymbol{x}\,|\,\boldsymbol{\theta})}{\partial A} = \frac{1}{\sigma^2} \sum_{n=1}^{N} x_n - A$$

$$\frac{\partial \log \mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)}{\partial \sigma^2} = -\frac{N}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N}\left(x_n - A\right)^2$$

$$\left(\frac{\partial^2 \log \mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)}{\partial \text{msup}} = -\frac{N}{\sigma^2}\right) \to -\frac{N}{\sigma^2}$$

$$\left(\frac{\partial^2 \log \mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)}{\partial A \partial \sigma^2} = -\left(\frac{1}{\sigma^4}\sum_{n=1}^{N} x_n - A\right)\right) \to 0$$

$$\left(\frac{\partial^2 \log \mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)}{\partial \text{msup}} = \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{n=1}^{N}\left(x_n - A\right)^2\right) \to -\frac{N}{2\sigma^4}$$

Which leads to

$$I(\boldsymbol{\theta}) = \begin{pmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{pmatrix}$$

$$\sigma\left(\widehat{A}\right)^2 \geq \frac{\sigma^2}{N}$$

$$\sigma\left(\widehat{\sigma^2}\right)^2 \geq \frac{2\sigma^4}{N}$$

Note that the CRLB for $\widehat{A}$ is the same whether or not σ^2 is known. This happens in this case due to the diagonal nature of the Fisher Information Matrix.
In general the Fisher Information Matrix is not diagonal and consequently the CRLBs will depend on other unknown parameters.

Glossary

idd

independent and identically distributed

Maximum Likelihood Estimation

The **maximum likelihood estimator** (MLE) is an alternative to the minimum variance unbiased estimator (MVUE). For many estimation problems, the MVUE does not exist. Moreover, when it does exist, there is no systematic procedure for finding it. In constrast, the MLE does not necessarily satisfy any optimality criterion, but it can almost always be computed, either through exact formulas or numerical techniques. For this reason, the MLE is one of the most common estimation procedures used in practice.

The MLE is an important type of estimator for the following reasons:

1. The MLE implements the likelihood principle.
2. MLEs are often simple and easy to compute.
3. MLEs have asymptotic optimality properties (consistency and efficiency).
4. MLEs are invariant under reparameterization.
5. If an efficient estimator exists, it is the MLE.
6. In signal detection with unknown parameters (composite hypothesis testing), MLEs are used in implementing the generalized likelihood ratio test (GLRT).

This module will discuss these properties in detail, with examples.

The Likelihood Principle

Supposed the data $\boldsymbol{X}$ is distributed according to the density or mass function $\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{\theta}\right)$. The **likelihood function** for $\boldsymbol{\theta}$ is defined by

$$\mathrm{l}\left(\theta|\boldsymbol{x}\right) \equiv \mathrm{p}\left(\boldsymbol{x}|\boldsymbol{\theta}\right)$$

At first glance, the likelihood function is nothing new - it is simply a way of rewriting the pdf/pmf of $\boldsymbol{X}$. The difference between the likelihood and the pdf or pmf is what is held fixed and what is allowed to vary. When we talk about the likelihood, we view the observation $\boldsymbol{x}$ as being fixed, and the parameter $\boldsymbol{\theta}$ as freely varying.

Note: It is tempting to view the likelihood function as a probability density for $\boldsymbol{\theta}$, and to think of $\mathrm{l}\left(\boldsymbol{\theta} \mid \boldsymbol{x}\right)$ as the conditional density of $\boldsymbol{\theta}$ given $\boldsymbol{x}$. This approach to parameter estimation is called **fiducial inference**, and is not accepted by most statisticians. One potential problem, for example, is that in many cases $\mathrm{l}\left(\boldsymbol{\theta} \mid \boldsymbol{x}\right)$ is not integrable ($\int \mathrm{l}\left(\boldsymbol{\theta} \mid \boldsymbol{x}\right) \mathrm{d}\, \boldsymbol{\theta} \rightarrow \infty$) and thus cannot be normalized. A more fundamental problem is that $\boldsymbol{\theta}$ is viewed as a fixed quantity, as opposed to random. Thus, it doesn't make sense to talk about its density. For the likelihood to be properly thought of as a density, a Bayesian approach is required.

The likelihood principle effectively states that all information we have about the unknown parameter $\boldsymbol{\theta}$ is contained in the likelihood function.
PrincipleLikelihood Principle

The information brought by an observation $\boldsymbol{x}$ about $\boldsymbol{\theta}$ is entirely contained in the likelihood function $\mathrm{p}\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)$. Moreover, if x_1 and x_2 are two observations depending on the same parameter $\boldsymbol{\theta}$, such that there exists a constant c satisfying $\mathrm{p}\left(x_1 \mid \boldsymbol{\theta}\right) = c\, \mathrm{p}\left(x_2 \mid \boldsymbol{\theta}\right)$ for every $\boldsymbol{\theta}$, then they bring the same information about $\boldsymbol{\theta}$ and must lead to identical estimators.
In the statement of the likelihood principle, it is **not** assumed that the two observations x_1 and x_2 are generated according to the same model, as long as the model is parameterized by $\boldsymbol{\theta}$.

Example:
Suppose a public health official conducts a survey to estimate $0 \leq \theta \leq 1$, the percentage of the population eating pizza at least once per week. As a result, the official found nine people who had eaten pizza in the last week, and three who had not. If no additional information is available regarding how the survey was implemented, then there are at least two probability models we can adopt.

1. The official surveyed 12 people, and 9 of them had eaten pizza in the last week. In this case, we observe $x_1 = 9$, where

$$x_1 \sim \text{Binomial}\,(12, \theta)$$

The density for x_1 is

$$\text{f}\,(x_1 | \boldsymbol{\theta}) = \binom{12}{x_1} \theta^{x_1} (1-\theta)^{12-x_1}$$

2. Another reasonable model is to assume that the official surveyed people **until** he found 3 non-pizza eaters. In this case, we observe $x_2 = 12$, where

$$x_2 \sim \text{NegativeBinomial}\,(3, 1-\theta)$$

The density for x_2 is

$$\text{g}\,(x_2 | \boldsymbol{\theta}) = \binom{x_2 - 1}{3 - 1} \theta^{x_2 - 3} (1-\theta)^3$$

The likelihoods for these two models are proportional:

$$\ell(\theta | x_1) \propto \ell(\theta | x_2) \propto \theta^9 (1-\theta)^3$$

Therefore, any estimator that adheres to the likelihood principle will produce the same estimate for θ, regardless of which of the two data-generation models is assumed.

The likelihood principle is widely accepted among statisticians. In the context of parameter estimation, any reasonable estimator should conform to the likelihood principle. As we will see, the maximum likelihood estimator does.

Note: While the likelihood principle itself is a fairly reasonable assumption, it can also be derived from two somewhat more intuitive assumptions known as the **sufficiency principle** and the **conditionality principle.** See Casella and Berger, Chapter 6.

The Maximum Likelihood Estimator

The **maximum likelihood estimator** $\widehat{\theta(\boldsymbol{x})}$ is defined by

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathrm{l} \left(\boldsymbol{\theta} | \boldsymbol{x}\right)$$

Intuitively, we are choosing $\boldsymbol{\theta}$ to maximize the probability of occurrence of the observation $\boldsymbol{x}$.

Note: It is possible that multiple parameter values maximize the likelihood for a given $\boldsymbol{x}$. In that case, any of these maximizers can be selected as the MLE. It is also possible that the likelihood may be **unbounded**, in which case the MLE does not exist.

The MLE rule is an implementation of the likelihood principle. If we have two observations whose likelihoods are proportional (they differ by a constant that does not depend on $\boldsymbol{\theta}$), then the value of $\boldsymbol{\theta}$ that maximizes one likelihood will also maximize the other. In other words, both likelihood functions lead to the same inference about θ, as required by the likelihood principle.

Understand that maximum likelihood is a **procedure**, not an optimality criterion. From the definition of the MLE, we have no idea how close it comes to the true parameter value relative to other estimators. In constrast, the MVUE is defined as the estimator that satisfies a certain optimality criterion. However, unlike the MLE, we have no clear produre to follow to compute the MVUE.

Computing the MLE

If the likelihood function is differentiable, then $\widehat{\boldsymbol{\theta}}$ is found by differentiating the likelihood (or log-likelihood), equating with zero, and solving:

$$\frac{\partial \log \mathrm{l}\left(\boldsymbol{\theta} | \boldsymbol{x}\right)}{\partial \boldsymbol{\theta}} = 0$$

If multiple solutions exist, then the MLE is the solution that maximizes $\log \mathrm{l}\left(\boldsymbol{\theta} | \boldsymbol{x}\right)$, that is, the **global** maximizer.

In certain cases, such as pdfs or pmfs with an esponential form, the MLE can be easily solved for. That is,

$$\frac{\partial \log \mathrm{l}\left(\boldsymbol{\theta} | \boldsymbol{x}\right)}{\partial \boldsymbol{\theta}} = 0$$

can be solved using calculus and standard linear algebra.

Example:
DC level in white Guassian noise
Suppose we observe an unknown amplitude in white Gaussian noise with unknown variance:

$$x_n = A + w_n$$

$n \in \{0, 1, \ldots, N-1\}$, where $w_n \sim \mathcal{N}\left(0, \sigma^2\right)$ are independent and identically distributed. We would like to estimate

$$\boldsymbol{\theta} = \begin{pmatrix} A \\ \sigma^2 \end{pmatrix}$$

by computing the MLE. Differentiating the log-likelihood gives

$$\frac{\partial \log \mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n - A$$

$$\frac{\partial \log \mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)}{\partial \sigma^2} = -\frac{N}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N} \left(x_n - A\right)^2$$

Equating with zero and solving gives us our MLEs:

$$\widehat{A} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

and

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N} \left(x_n - \widehat{A}\right)^2$$

Note: $\widehat{\sigma^2}$ is biased!

As an exercise, try the following problem:

Exercise:

Problem:

Suppose we observe a random sample $\boldsymbol{x} = \left(x_1 \ldots x_N\right)^T$ of Poisson measurements with intensity λ: $\mathrm{Pr}[x_i = n] = e^{-\lambda}\frac{\lambda^n}{n!}$, $n \in \{0, 1, 2, \ldots\}$. Find the MLE for λ.

Unfortunately, this approach is only feasible for the most elementary pdfs and pmfs. In general, we may have to resort to more advanced numerical maximization techniques:

1. **Newton-Raphson** iteration
2. Iteration by the **Scoring Method**
3. **Expectation-Maximization Algorithm**

All of these are iterative techniques which posit some initial guess at the MLE, and then incrementally update that guess. The iteration procedes until a local maximum of the likelihood is attained, although in the case of the first two methods, such convergence is not guaranteed. The EM algorithm has the advantage that the likelihood is always increased at each iteration, and so convergence to at least a local maximum is guaranteed (assuming a bounded likelihood). For each algorithm, the final estimate is highly dependent on the initial guess, and so it is customary to try several different starting values. For details on these algorithms, see Kay, Vol. I.

Asymptotic Properties of the MLE

Let $\boldsymbol{x} = (x_1 \ldots x_N)^T$ denote an IID sample of size N, and each sample is distributed according to $\mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)$. Let $\hat{\theta}_N$ denote the MLE based on a sample $\boldsymbol{x}$.

Asymptotic Properties of MLE

If the likelihood $\ell(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = \mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)$ satisfies certain "regularity" conditions[footnote], then the MLE $\hat{\theta}_N$ is **consistent**, and moreover, $\hat{\theta}_N$ converges in probability to $\widehat{\boldsymbol{\theta}}$, where

$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \left(I(\boldsymbol{\theta})\right)^{-1}\right)$$

where $I(\boldsymbol{\theta})$ is the **Fisher Information matrix** evaluated at the true value of $\boldsymbol{\theta}$.

The regularity conditions are essentially the same as those assumed for the Cramer-Rao lower bound: the log-likelihood must be twice differentiable, and the expected value of the first derivative of the log-likelihood must be zero.

Since the mean of the MLE tends to the true parameter value, we say the MLE is **asymptotically unbiased**. Since the covariance tends to the inverse Fisher information matrix, we say the MLE is **asymptotically efficient**.

In general, the rate at which the mean-squared error converges to zero is not known. It is possible that for small sample sizes, some other estimator may have a smaller MSE.The proof of consistency is an application of the weak law of large numbers. Derivation of the asymptotic distribution relies on the central limit theorem. The theorem is also true in more general settings (e.g., dependent samples). See, Kay, Vol. I, Ch. 7 for further discussion.

The MLE and Efficiency

In some cases, the MLE is efficient, not just asymptotically efficient. In fact, when an efficient estimator exists, it must be the MLE, as described by the following result:

If $\widehat{\boldsymbol{\theta}}$ is an efficient estimator, and the Fisher information matrix $I(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta}$, then $\widehat{\boldsymbol{\theta}}$ maximizes the likelihood.

Recall the $\widehat{\boldsymbol{\theta}}$ is efficient (meaning it is unbiased and achieves the Cramer-Rao lower bound) if and only if

$$\frac{\partial \ln(\mathrm{p}\ (\boldsymbol{x}\,|\,\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}} = I(\boldsymbol{\theta})\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$$

for all $\boldsymbol{\theta}$ and $\boldsymbol{x}$. Since $\widehat{\boldsymbol{\theta}}$ is assumed to be efficient, this equation holds, and in particular it holds when $\boldsymbol{\theta} = \widehat{\theta(\boldsymbol{x})}$. But then the derivative of the log-likelihood is zero at $\boldsymbol{\theta} = \widehat{\theta(\boldsymbol{x})}$. Thus, $\widehat{\boldsymbol{\theta}}$ is a critical point of the likelihood. Since the Fisher information matrix, which is the negative of the matrix of second order derivatives of the log-likelihood, is positive definite, $\widehat{\boldsymbol{\theta}}$ must be a maximum of the likelihood.

An important case where this happens is described in the following subsection.

Optimality of MLE for Linear Statistical Model

If the observed data $\boldsymbol{x}$ are described by

$$\boldsymbol{x} = H\boldsymbol{\theta} + \boldsymbol{w}$$

where H is $N \times p$ with full rank, $\boldsymbol{\theta}$ is $p \times 1$, and $\boldsymbol{w} \sim \mathscr{N}(\mathbf{0}, C)$, then the MLE of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \left(H^T C^{-1} H\right)^{-1} H^T C^{-1} \boldsymbol{x}$$

This can be established in two ways. The first is to compute the CRLB for $\boldsymbol{\theta}$. It turns out that the condition for equality in the bound is satisfied, and $\widehat{\boldsymbol{\theta}}$ can be read off from that condition.

The second way is to maximize the likelihood directly. Equivalently, we must minimize

$$(\boldsymbol{x} - H\boldsymbol{\theta})^T C^{-1} (\boldsymbol{x} - H\boldsymbol{\theta})$$

with respect to $\boldsymbol{\theta}$. Since C^{-1} is positive definite, we can write $C^{-1} = U^T \mathit{\Lambda} U = D^T D$, where $D = \mathit{\Lambda}^{\frac{1}{2}} U$, where U is an orthogonal matrix whose columns are eigenvectors of C^{-1}, and $\mathit{\Lambda}$ is a diagonal matrix with positive diagonal entries. Thus, we must minimize

$$(D\boldsymbol{x} - DH\boldsymbol{\theta})^T (D\boldsymbol{x} - DH\boldsymbol{\theta})$$

But this is a linear least squares problem, so the solution is given by the pseudoinverse of DH:

Equation:

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \left((DH)^T (DH)\right)^{-1} (DH)^T (D\boldsymbol{x}) \\ &= \left(H^T C^{-1} H\right)^{-1} H^T C^{-1} \boldsymbol{x}\end{aligned}$$

Exercise:

Problem:

Consider $X_1, \ldots, X_N \sim \mathcal{N}\left(\boldsymbol{s}, \sigma^2 I\right)$, where $\boldsymbol{s}$ is a $p \times 1$ unknown signal, and σ^2 is known. Express the data in the linear model and find the MLE $\widehat{\boldsymbol{s}}$ for the signal.

Invariance of MLE

Suppose we wish to estimate the function $\boldsymbol{w} = W(\boldsymbol{\theta})$ and not $\boldsymbol{\theta}$ itself. To use the maximum likelihood approach for estimating $\boldsymbol{w}$, we need an expression for the likelihood $\ell(\boldsymbol{w} | \boldsymbol{x})$ =p $(\boldsymbol{x} | \boldsymbol{w})$. In other words, we would need to be able to parameterize the distribution of the data by $\boldsymbol{w}$. If W is not a one-to-one function, however, this may not be possible. Therefore, we define the **induced** likelihood

$$\ell(\boldsymbol{w} | \boldsymbol{x}) = \text{max}_{\boldsymbol{\theta}} \ \{\boldsymbol{\theta}, W(\boldsymbol{\theta}) = \boldsymbol{w}\} \ell(\boldsymbol{\theta} | \boldsymbol{x})$$

The MLE $\widehat{\boldsymbol{w}}$ is defined to be the value of $\boldsymbol{w}$ that maximizes the induced likelihood. With this definition, the following invariance principle is immediate.

Let $\widehat{\boldsymbol{\theta}}$ denote the MLE of $\boldsymbol{\theta}$. Then $\widehat{\boldsymbol{w}} = W\left(\widehat{\boldsymbol{\theta}}\right)$ is the MLE of $\boldsymbol{w} = W(\boldsymbol{\theta})$.

The proof follows directly from the definitions of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{w}}$. As an exercise, work through the logical steps of the proof on your own.

Example:
Let $\boldsymbol{x} = (x_1 \ldots x_N)^T$ where

$$x_i \sim \mathrm{Poisson}\,(\lambda)$$

Given $\boldsymbol{x}$, find the MLE of the probability that $x \sim \mathrm{Poisson}\,(\lambda)$ exceeds the mean λ.

$$W(\lambda) = \Pr[x > \lambda] = \sum_{n=\lfloor \lambda+1 \rfloor}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!}$$

where $\lfloor z \rfloor =$ largest integer $\leq z$. The MLE of w is

$$\widehat{w} = \sum_{n=\left\lfloor \widehat{\lambda}+1 \right\rfloor}^{\infty} e^{-\widehat{\lambda}} \frac{\left(\widehat{\lambda}\right)^n}{n!}$$

where $\widehat{\lambda}$ is the MLE of λ:

$$\widehat{\lambda} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

Be aware that the MLE of a **transformed** parameter does not necessarily satisfy the asymptotic properties discussed earlier.

Exercise:

Problem:

Consider observations x_1,…,x_N, where x_i is a p-dimensional vector of the form $x_i = \boldsymbol{s} + w_i$ where $\boldsymbol{s}$ is an unknown signal and w_i are independent realizations of white Gaussian noise:

$$w_i \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I_{p \times p}\right)$$

Find the maximum likelihood estimate of the energy $E = \boldsymbol{s}^T \boldsymbol{s}$ of the unknown signal.

Summary of MLE

The likelihood principle states that information brought by an observation $\boldsymbol{x}$ about $\boldsymbol{\theta}$ is entirely contained in the likelihood function $\mathrm{p}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)$. The maximum likelihood estimator is **one** effective implementation of the likelihood principle. In some cases, the MLE can be computed exactly, using calculus and linear algebra, but at other times iterative numerical algorithms are needed. The MLE has several desireable properties:

- It is consistent and asymptotically efficient (as $N \to \infty$ we are doing as well as MVUE).
- When an efficient estimator exists, it is the MLE.
- The MLE is invariant to reparameterization.

Bayesian Estimation

We are interested in estimating $\boldsymbol{\theta}$ given the observation $\boldsymbol{x}$. Naturally then, any estimation strategy will be based on the posterior distribution $\mathrm{p}\ (\boldsymbol{\theta}|\boldsymbol{x})$. Furthermore, we need a criterion for assessing the quality of potential estimators.

Loss

The quality of an estimate $\hat{\theta}$ is measured by a real-valued **loss function**: $L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)$. For example, squared error or quadratic loss is simply $L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)=\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)^T\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)$

Expected Loss

Posterior Expected Loss:

$$E\left[L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)|\boldsymbol{x}\right]=\int L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)\mathrm{p}\ (\boldsymbol{\theta}|\boldsymbol{x})\ \mathrm{d}\ \boldsymbol{\theta}$$

Bayes Risk:
Equation:

$$\begin{aligned}E\left[L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)\right] &= \int\int L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)\mathrm{p}\ (\boldsymbol{\theta}|\boldsymbol{x})\ \mathrm{p}\ (\boldsymbol{x})\ \mathrm{d}\ \boldsymbol{\theta}\ \mathrm{d}\ \boldsymbol{x}\\ &= \int\int L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)\mathrm{p}\ (\boldsymbol{x}|\boldsymbol{\theta})\ \mathrm{p}\ (\boldsymbol{\theta})\ \mathrm{d}\ \boldsymbol{x}\ \mathrm{d}\ \boldsymbol{\theta}\\ &= E\left[E\left[L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)|\boldsymbol{x}\right]\right]\end{aligned}$$

The "best" or optimal estimator given the data $\boldsymbol{x}$ and under a specified loss is given by

$$\widehat{\boldsymbol{\theta}}=\underset{\boldsymbol{\theta}}{\mathrm{argmin}}\ E\left[L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)|\boldsymbol{x}\right]$$

Example:
Bayes MSE

$$\mathrm{BMSE}\left(\widehat{\boldsymbol{\theta}}\right) \equiv \int \int \left(\theta - \widehat{\boldsymbol{\theta}}\right)^2 \mathrm{p}\left(\theta | \boldsymbol{x}\right) \mathrm{d}\,\theta\, \mathrm{p}\left(\boldsymbol{x}\right) \mathrm{d}\,\boldsymbol{x}$$

Since $\mathrm{p}\left(\boldsymbol{x}\right) \geq 0$ for every $\boldsymbol{x}$, minimizing the inner integral $\int \left(\theta - E[\boldsymbol{\theta}]\right)^2 \mathrm{p}\left(\theta | \boldsymbol{x}\right) \mathrm{d}\,\theta = E\left[L\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) | \boldsymbol{x}\right]$ (where $E\left[L\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) | \boldsymbol{x}\right]$ is the posterior expected loss) for each $\boldsymbol{x}$, minimizes the overall BMSE.

Equation:

$$\begin{aligned} \frac{\partial \int \left(\theta - \widehat{\boldsymbol{\theta}}\right)^2 \mathrm{p}(\theta | \boldsymbol{x}) \mathrm{d}\theta}{\partial \widehat{\boldsymbol{\theta}}} &= \int \frac{\partial \left(\left(\theta - \widehat{\boldsymbol{\theta}}\right)^2 \mathrm{p}(\theta | \boldsymbol{x})\right)}{\partial \widehat{\boldsymbol{\theta}}} \mathrm{d}\,\theta \\ &= -2 \int \left(\theta - \widehat{\boldsymbol{\theta}}\right) \mathrm{p}\left(\theta | \boldsymbol{x}\right) \mathrm{d}\,\theta \end{aligned}$$

Equating this to zero produces

$$\widehat{\boldsymbol{\theta}} = \int \theta\, \mathrm{p}\left(\theta | \boldsymbol{x}\right) \mathrm{d}\,\theta \equiv E[\theta | \boldsymbol{x}]$$

The conditional mean (also called **posterior mean**) of θ given $\boldsymbol{x}$!

Example:

$$\forall n, n \in \{1, \ldots, N\} : \left(x_n = A + W_n\right)$$

$$W_n \sim \mathcal{N}\left(0, \sigma^2\right)$$

prior for unknown parameter A:

$$\mathrm{p}\left(a\right) = U(-A_0, A_0)$$

$$\mathrm{p}\left(\boldsymbol{x} | A\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} e^{\frac{-1}{2\sigma^2} \sum_{n=1}^{N} \left(x_n - A\right)^2}$$

$$\mathrm{p}\left(A|\boldsymbol{x}\right) = \begin{cases} \dfrac{\frac{1}{2A_0\left(2\pi\sigma^2\right)^{\frac{N}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(xn-A)^2}}{\int_{-A_0}^{A_0} \frac{1}{2A_0\left(2\pi\sigma^2\right)^{\frac{N}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(xn-A)^2} \mathrm{d}A} & \text{if} \quad |A| \leq A_0 \\ 0 \ \text{if} \quad |A| > A_0 \end{cases}$$

Minimum Bayes MSE Estimator:
Equation:

$$\begin{aligned} \widehat{A} &= E[A|\boldsymbol{x}] \\ &= \int_{-\infty}^{\infty} a \ \mathrm{p}\left(A|\boldsymbol{x}\right) \mathrm{d}\ A \\ &= \frac{\int_{-A_0}^{A_0} A \frac{1}{2A_0\left(2\pi\sigma^2\right)^{\frac{N}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(xn-A)^2} \mathrm{d}A}{\int_{-A_0}^{A_0} \frac{1}{2A_0\left(2\pi\sigma^2\right)^{\frac{N}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(xn-A)^2} \mathrm{d}A} \end{aligned}$$

Notes

1. No closed-form estimator
2. As $A_0 \to \infty$, $\widehat{A} \to \frac{1}{N}\sum_{n=1}^{N} x_n$
3. For smaller A_0, truncated integral produces an $\widehat{A}$ that is a function of $\boldsymbol{x}$, σ^2, and A_0
4. As N increases, $\frac{\sigma^2}{N}$ decreases and posterior $\mathrm{p}\left(A|\boldsymbol{x}\right)$ becomes tightly clustered about $\frac{1}{N}\sum x_n$. This implies $\widehat{A} \to \frac{1}{N}\sum_n x_n$ as $n \to \infty$ (the data "swamps out" the prior)

Other Common Loss Functions

Absolute Error Loss

(Laplace, 1773)

$$L\left(\theta,\hat{\theta}\right)=\left|\theta-\hat{\theta}\right|$$

Equation:

$$\begin{aligned}E\left[L\left(\theta,\hat{\theta}\right)|\boldsymbol{x}\right] &= \int_{-\infty}^{\infty}\left|\theta-\hat{\theta}\right|\mathrm{p}\ (\theta|\boldsymbol{x})\ \mathrm{d}\ \theta\\ &= \int_{-\infty}^{\hat{\theta}}\left(\hat{\theta}-\theta\right)\mathrm{p}\ (\theta|\boldsymbol{x})\ \mathrm{d}\ \theta+\int_{\hat{\theta}}^{\infty}\left(\theta-\hat{\theta}\right)\mathrm{p}\ (\theta|\boldsymbol{x})\ \mathrm{d}\ \theta\end{aligned}$$

Using integration-by-parts it can be shown that

$$\int_{-\infty}^{\hat{\theta}}\left(\hat{\theta}-\theta\right)\mathrm{p}\ (\theta|\boldsymbol{x})\ \mathrm{d}\ \theta=\int_{-\infty}^{\hat{\theta}}\mathrm{P}\ (\theta<y|\boldsymbol{x})\ \mathrm{d}\ y$$

$$\int_{\hat{\theta}}^{\infty}\left(\theta-\hat{\theta}\right)\mathrm{p}\ (\theta|\boldsymbol{x})\ \mathrm{d}\ \theta=\int_{\hat{\theta}}^{\infty}\mathrm{P}\ (\theta>y|\boldsymbol{x})\ \mathrm{d}\ y$$

where $\mathrm{P}\ (\theta<y|\boldsymbol{x})$ and $\mathrm{P}\ (\theta>y|\boldsymbol{x})$ are a cumulative distributions. So,

$$E\left[L\left(\theta,\hat{\theta}\right)|\boldsymbol{x}\right]=\int_{-\infty}^{\hat{\theta}}\mathrm{P}\ (\theta<y|\boldsymbol{x})\ \mathrm{d}\ y+\int_{\hat{\theta}}^{\infty}\mathrm{P}\ (\theta>y|\boldsymbol{x})\ \mathrm{d}\ y$$

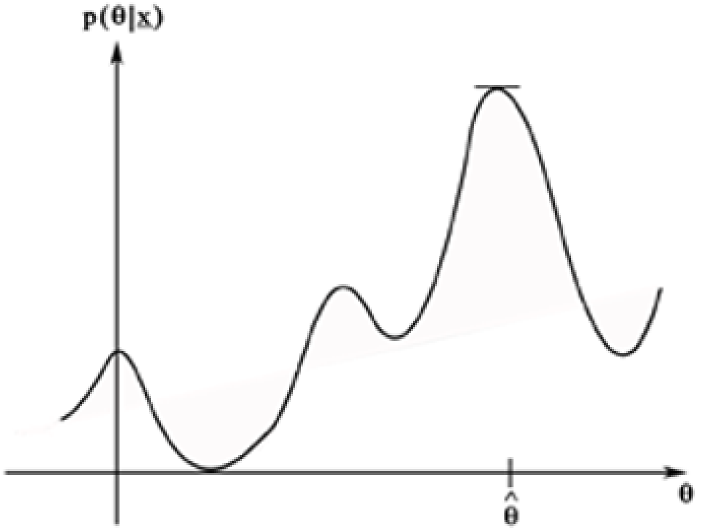
Take the derivative with respect to $\hat{\theta}$ implies $\mathrm{P}\ \left(\theta<\hat{\theta}|\boldsymbol{x}\right)$ $=\mathrm{P}\ \left(\theta>\hat{\theta}|\boldsymbol{x}\right)$ which implies that the optimal $\hat{\theta}$ under absolute error loss is **posterior median**.

'0-1' Loss

$$L\left(\theta,\hat{\theta}\right)=\begin{cases}0\ \text{ if }\ \hat{\theta}=\theta\\ 1\ \text{ if }\ \hat{\theta}\neq\theta\end{cases}=I_{\left\{\hat{\theta}\neq\theta\right\}}$$

$$E\left[L\left(\theta,\hat{\theta}\right)|\boldsymbol{x}\right]=E\left[I_{\left\{\hat{\theta}\neq\theta\right\}}|\boldsymbol{x}\right]=\Pr\left[\hat{\theta}\neq\theta\ \middle|\ \boldsymbol{x}\right]$$

which is the probability that $\hat{\theta} \neq \theta$ given $\boldsymbol{x}$. To minimize '0-1' loss we must choose $\hat{\theta}$ to be the value of θ with the highest posterior probability, which implies $\hat{\theta} \neq \theta$ with the smallest probability.



The optimal estimator $\hat{\theta}$ under '0-1' loss is the **maximum a posteriori** (MAP) estimator--the value of θ where $\mathrm{p}\left(\theta \mid \boldsymbol{x}\right)$ is maximized.

Wiener Filtering and the DFT

Connecting the Vector Space and Classical Wiener Filters

Suppose we observe

$$\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{w}$$

which are all $N \times 1$ vectors and where $\boldsymbol{w} \sim \mathcal{N}\left(0, \sigma^2 I\right)$. Given $\boldsymbol{x}$ we wish to estimate $\boldsymbol{y}$. Think of $\boldsymbol{y}$ as a signal in additive white noise $\boldsymbol{w}$. $\boldsymbol{x}$ is a noisy observation of the signal.

Taking a Bayesian approach, put a prior on the signal $\boldsymbol{y}$:

$$\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, R_{\mathrm{yy}})$$

which is independent of noise $\boldsymbol{w}$. The minimum MSE (MMSE) estimator is

$$\widehat{\boldsymbol{y}} = R_{\mathrm{yx}} {R_{\mathrm{xx}}}^{-1} \boldsymbol{x}$$

Under the modeling assumptions above
Equation:

$$\begin{array}{rcl} R_{\mathrm{yx}} & = & E\left[\boldsymbol{y}(\boldsymbol{y}+\boldsymbol{w})^T\right] \\ & = & E\left[\boldsymbol{y}\boldsymbol{y}^T\right] + E\left[\boldsymbol{y}\boldsymbol{w}^T\right] \\ & = & E\left[\boldsymbol{y}\boldsymbol{y}^T\right] \\ & = & R_{\mathrm{yy}} \end{array}$$

since $E\left[\boldsymbol{y}\boldsymbol{w}^T\right] = 0$ and since $\boldsymbol{y}$ and $\boldsymbol{w}$ are zero-mean and independent.
Equation:

$$\begin{aligned} R_{\mathrm{xx}} &= E\left[\boldsymbol{x}\boldsymbol{x}^T\right] \\ &= E\left[(\boldsymbol{y}+\boldsymbol{w})(\boldsymbol{y}+\boldsymbol{w})^T\right] \\ &= E\left[\boldsymbol{y}\boldsymbol{y}^T\right] + E\left[\boldsymbol{y}\boldsymbol{w}^T\right] + E\left[\boldsymbol{w}\boldsymbol{y}^T\right] + E\left[\boldsymbol{w}\boldsymbol{w}^T\right] \\ &= R_{\mathrm{yy}} + R_{\mathrm{ww}} \end{aligned}$$

since $E\left[\boldsymbol{w}\boldsymbol{w}^T\right] = R_{\mathrm{ww}}$. Hence

$$\widehat{\boldsymbol{y}} = R_{\mathrm{yy}}(R_{\mathrm{yy}} + R_{\mathrm{ww}})^{-1}\boldsymbol{x} = H_{\mathrm{opt}}\boldsymbol{x}$$

Where H_{opt} is the Wiener filter. Recall the frequency domain case

$$H_{\mathrm{opt}}(f) = \frac{S_{\mathrm{yy}}(f)}{S_{\mathrm{yy}}(f) + S_{\mathrm{ww}}(f)}$$

Now let's look at an actual problem scenario. Suppose that we know a priori that the signal $\boldsymbol{y}$ is **smooth** or **lowpass**. We can incorporate this prior knowledge by carefully choosing the prior covariance R_{yy}.

Recall the DFT

$$\forall k, k = 0, \ldots, N-1 : \left(\mathscr{Y}_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N} Y_k e^{-\left(i2\pi\frac{kn}{N}\right)}\right)$$

or in vector notation

$$\forall k, k = 0, \ldots, N-1 : \left(\mathscr{Y}_k = \langle \boldsymbol{y}, u_k \rangle\right)$$

where $u_k = \frac{\begin{pmatrix} 1 & e^{i2\pi\frac{k}{N}} & e^{i2\pi\frac{2k}{N}} & \dots & e^{i2\pi\frac{(N-1)k}{N}} \end{pmatrix}^{\mathrm{H}}}{\sqrt{N}}$ (H dehotes Hermitian transpose)

Note: $\langle u_k, u_k \rangle = {u_k}^{\mathrm{H}} u_k = 1, \forall k \wedge l, k \neq l : \left(\langle u_k, u_l \rangle = {u_k}^{\mathrm{H}} u_l = 0\right)$, i.e., $\forall k, k = 0, \ldots, N-1 : (\{u_k\})$ is an orthonormal basis.

The vector u_k spans the subspace corresponding to a frequency band centered at frequency $f_k = \frac{2\pi k}{N}$ ("digital" frequency on $[0, 1]$). If we know that $\boldsymbol{y}$ is lowpass, then

$$E\left[(\| \langle \boldsymbol{y}, u_k \rangle \|)^2\right] = E\left[(\| \mathscr{Y}_k \|)^2\right]$$

should be relatively small (compared to $E\left[(\| \langle \boldsymbol{y}, u_0 \rangle \|)^2\right]$) for high frequencies.

Let

$$\sigma_k{}^2 = E\left[(\| \langle \boldsymbol{y}, u_k \rangle \|)^2\right]$$

A lowpass model implies $\sigma_0{}^2 > \sigma_1{}^2 > \ldots > \sigma_{\frac{N}{2}}{}^2$, assuming N even, and conjugate symmetry implies $\forall j, j = 1, \ldots, \frac{N}{2} : \left(\sigma_{N-j}{}^2 = \sigma_j{}^2\right)$ Furthermore, let's model the DFT coefficients as zero-mean and independent

$$E[\mathscr{Y}_k] = 0$$

$$E\left[\mathscr{Y}_k \overline{\mathscr{Y}_l}\right] = \begin{cases} \sigma_k{}^2 \text{ if } l = k \\ 0 \text{ if } l \neq k \end{cases}$$

This completely specifies our prior

$$\boldsymbol{y} \sim \mathscr{N}(\boldsymbol{0}, R_{\mathrm{yy}})$$

$$R_{\mathrm{yy}} = U D \overline{U}^T$$

where

$$D = \begin{pmatrix} {\sigma_0}^2 & 0 & \dots & 0 \\ 0 & {\sigma_1}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & {\sigma_{N-1}}^2 \end{pmatrix}$$

and

$$U = \begin{pmatrix} u_0 & u_1 & \dots & u_{N-1} \end{pmatrix}$$

Note:

$$\mathscr{Y} = U^{\mathrm{H}} \boldsymbol{y}$$

is the DFT and

$$\boldsymbol{y} = U \mathscr{Y}$$

is the inverse DFT.

With this prior on $\boldsymbol{y}$ the Wiener filter is

$$\widehat{\boldsymbol{y}} = UDU^{\mathrm{H}} \left(UDU^{\mathrm{H}} + \sigma^2 I \right)^{-1} \boldsymbol{x}$$

Since U is a unitary matrix $UU^{\mathrm{H}} = I$ and therefore
Equation:

$$\begin{aligned} \widehat{\boldsymbol{y}} &= UDU^{\mathrm{H}} \left(U \left(D + \sigma^2 I \right) U^{\mathrm{H}} \right)^{-1} \boldsymbol{x} \\ &= UDU^{\mathrm{H}} U \left(D + \sigma^2 I \right)^{-1} U^{\mathrm{H}} \boldsymbol{x} \\ &= UD \left(D + \sigma^2 I \right)^{-1} U^{\mathrm{H}} \boldsymbol{x} \end{aligned}$$

[footnote] Now take the DFT of both sides

$$\widehat{\mathscr{Y}} = U^{\mathrm{H}}\widehat{\boldsymbol{y}} = D\left(D + \sigma^2 I\right)^{-1}\mathscr{X}$$

where $\mathscr{X} = U^{\mathrm{H}}\boldsymbol{x}$ and is the DFT of $\boldsymbol{x}$. Both D and $D + \sigma^2 I$ are diagonal so

$$\widehat{\mathscr{Y}_k} = \frac{d_{k,k}}{d_{k,k} + \sigma^2}\mathscr{X}_k = \frac{{\sigma_k}^2}{{\sigma_k}^2 + \sigma^2}\mathscr{X}_k$$

Hence the Wiener filter is a frequency (DFT) domain filter

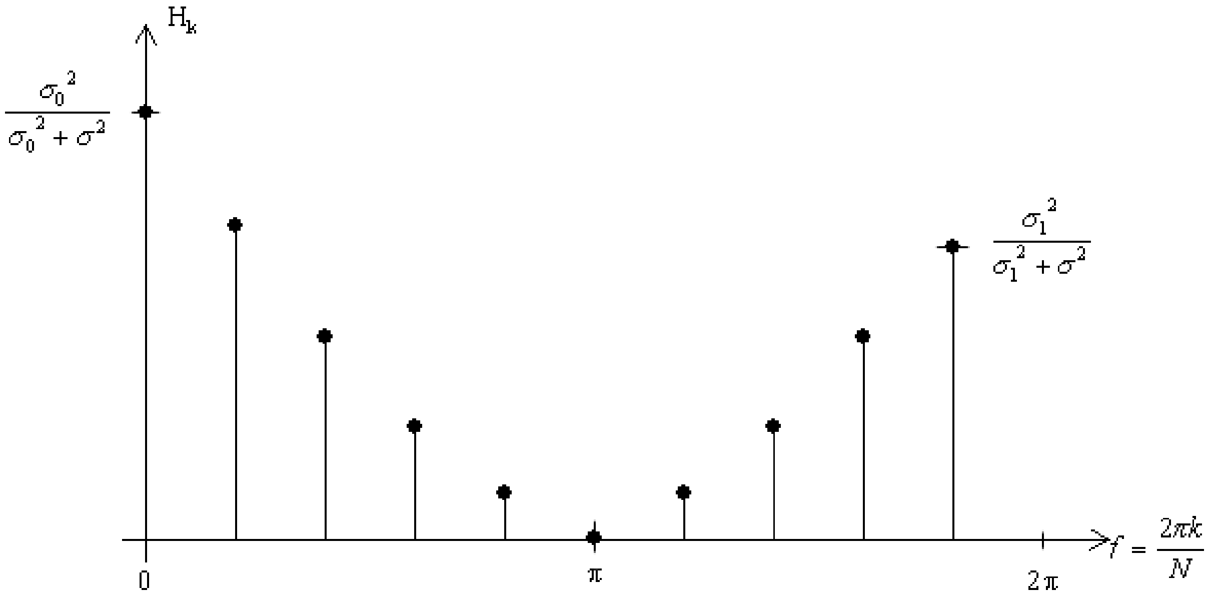
$$\widehat{\mathscr{Y}_k} = H_k\mathscr{X}_k$$

where $\mathscr{X}_k$ is the k^{th} DFT coefficient of $\boldsymbol{x}$ and the filter response at digital frequency $\frac{2\pi k}{N}$ is

$$H_k = \frac{{\sigma_k}^2}{{\sigma_k}^2 + \sigma^2}$$

Assuming ${\sigma_0}^2 > {\sigma_1}^2 > \ldots > {\sigma_{\frac{N}{2}}}^2$ and $\forall j, j = 1, \ldots, \frac{N}{2} : \left({\sigma_{N-j}}^2 = {\sigma_j}^2\right)$. The filter's response is a **digital lowpass filter**!

A Digital Lowpass Filter!



If A, B, C are all invertible, compatible matrices, then $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$. $U^{-1} = U^{\mathrm{H}}$, $\left(U^{\mathrm{H}}\right)^{-1} = U$.

Summary of Wiener Filter

Problem: Observe $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{w}$

Recover/estimate signal $\boldsymbol{y}$.

Classical Wiener Filter (continuous-time):

$$H(\omega) = \frac{S_{\mathrm{yy}}(\omega)}{S_{\mathrm{yy}}(\omega) + S_{\mathrm{ww}}(\omega)}$$

where $y(t)$ and $w(t)$ are stationary processes.

Vector Space Wiener Filter:

$$H = R_{\mathrm{yy}}(R_{\mathrm{yy}} + R_{\mathrm{ww}})^{-1}$$

Wiener Filter and DFT: ($R_{\mathrm{ww}} = \sigma^2 I$). If $R_{\mathrm{yy}} = UDU^{\mathrm{H}}$, where U is DFT, then H is a discrete-time filter whose DFT is given by
Equation:

$$\begin{aligned} H_k &= \sum_{n=0}^{N-1} h_n e^{i2\pi \frac{k}{N}} \\ &= \frac{d_{k,k}}{d_{k,k}+\sigma^2} \end{aligned}$$

Here, $d_{k,k}$ plays the same role as $S_{\mathrm{yy}}(\omega)$.

Kalman Filters

The Kalman filter is an important generalization of the Wiener filter. Unlike Wiener filters, which are designed under the assumption that the signal and noise are **stationary**, the Kalman filter has the ability to **adapt** itself to non-stationary environments.

The Kalman filter can be viewed as a **sequential** minimum MSE estimator of a signal in additive noise. If the signal and noise are jointly Gaussian, the then Kalman filter is optimal in a minimum MSE sense (minimizes expected quadratic loss).

If the signal and/or noise are non-Gaussian, then the Kalman filter is the best linear estimator (linear estimator that minimizes MSE among all possible linear estimators).

Dynamical Signal Models

Recall the simple DC signal estimation problem.
Equation:

$$\forall n, n = \{0, \ldots, N-1\} : (x_n = A + w_n)$$

Where A is the unknown DC level and w_n is the white Gaussian noise. A could represent the voltage of a DC power supply. We know how to find several good estimators of A given the measurements $\{x_0, \ldots, x_{N-1}\}$.

In practical situations this model may be too simplistic. the load on the power supply may charge over time and there will be other variations due to temperature and component aging.

To account for these variations we can employ a more accurate measurement model:
Equation:

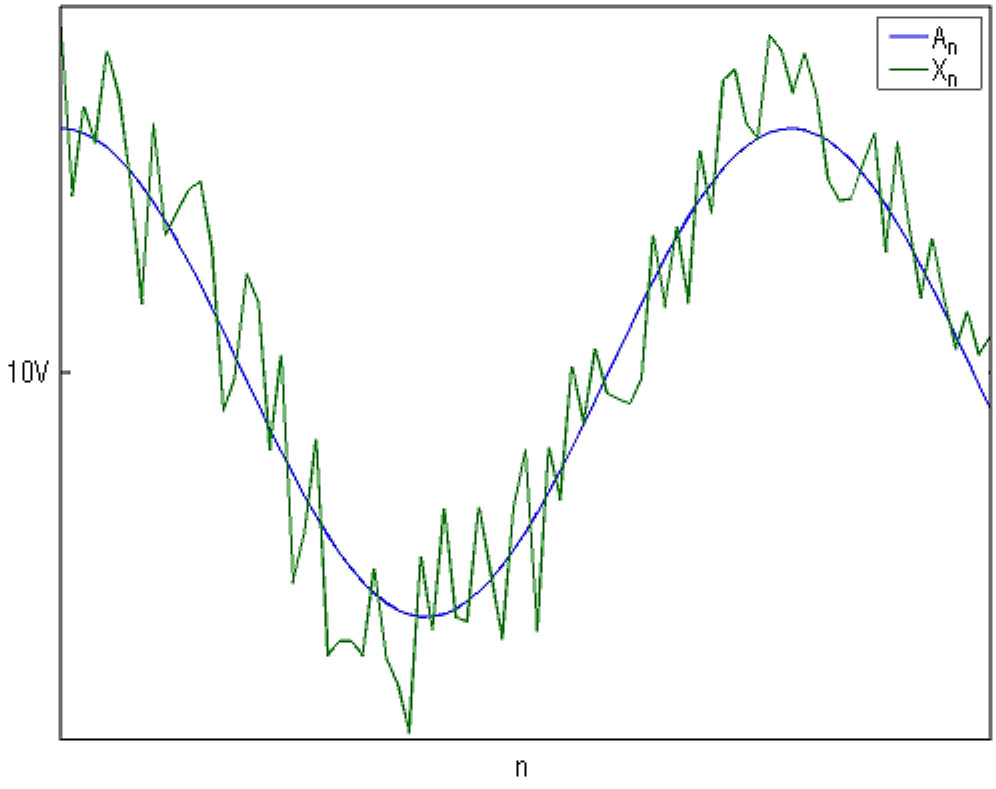
$$\forall n, n = \{0, \ldots, N-1\} : (x_n = A_n + w_n)$$

where the voltage A_n is the **true** voltage at time n.

Now the estimation problem is significantly more complicated since we must estimate $\{A_0, ..., A_{N-1}\}$. Suppose that the true voltage A_n does not vary too rapidly over time. Then successive samples of A_n will not be too different, suggesting that the voltage signal displays a high degree of **correlation**.

This reasoning suggests that it may be reasonable to regard the sequence $\{A_0, ..., A_{N-1}\}$, as a realization of a correlated (not white) random process. Adopting a random process model for A_n allows us to pursue a Bayesian approach to the estimation problem ([link]).

Voltage Varying Over Time



Using the model in [link], it is easy to verify that the maximum likelihood and MVUB esitmators are given by

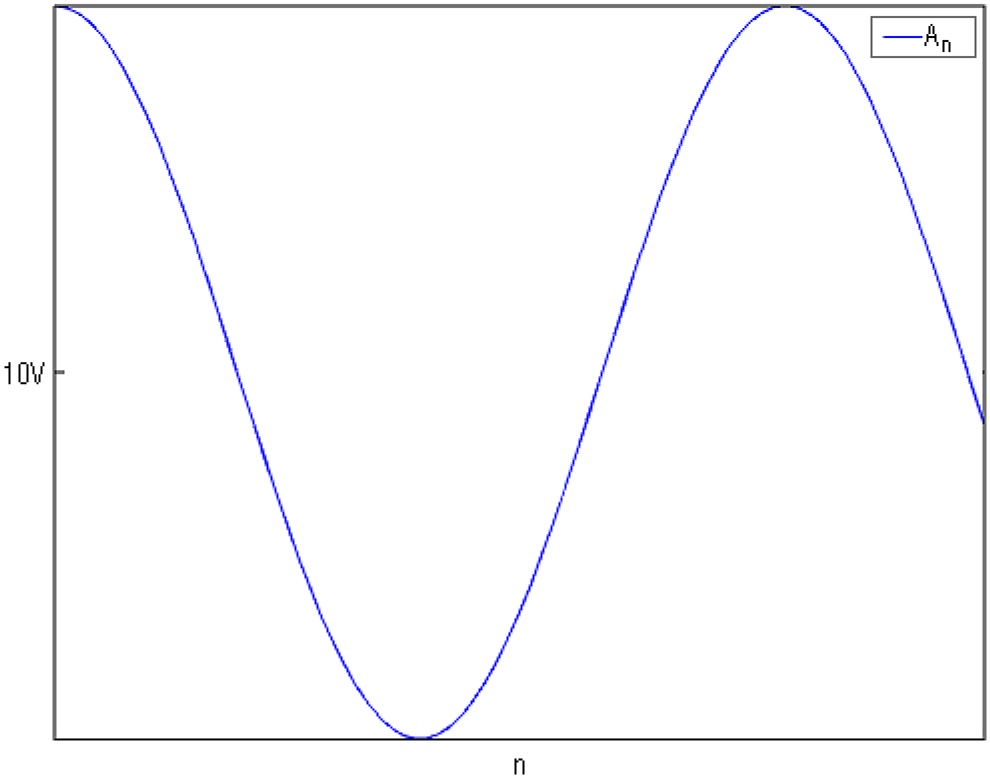
Equation:

$$\widehat{A}_n = x_n$$

Our estimate is simply the noisy measurements! No averaging takes place, so there is no noise reduction.

Let's look at the example again, [link].

True Voltage Varying Over Time



The voltage A_n is varying about an average value of 10V. Assume this average value is known and write

Equation:

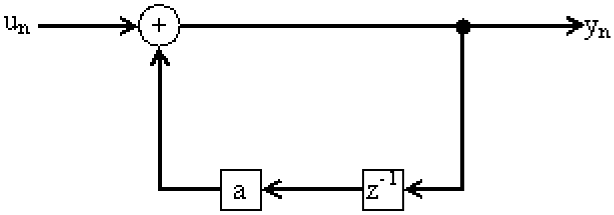
$$A_n = 10 + y_n$$

Where y_n is a zero-mean random process. Now a simple model for y_n which allows us to specify the correlation between samples is the **first-**

order Gauss-Markov prcoess model:
Equation:

$$\forall n, n = \{1, 2, \dots\} : (y_n = a y_{n-1} + u_n)$$

Where $u_n \sim \mathcal{N}\left(0, {\sigma_u}^2\right)$ iid (white Gaussian noise process). To initialize the process we take y_0 to be the realization of a Gaussian random variable: $y_0 \sim \mathcal{N}\left(0, {\sigma_y}^2\right)$. u_n is called the **driving** or **excitation** noise. The model in [link] is called the **dynamical** or **state** model. The current output y_n depends only on the **state** of the system at the previous time, or y_{n-1}, and the current input u_n ([link]).



$$y_1 = a y_0 + u_0$$

Equation:

$$\begin{aligned} y_2 &= a y_1 + u_1 \\ &= a\left(a y_0 + u_0\right) + u_1 \\ &= a^2 y_0 + a u_1 + u_2 \end{aligned}$$

$$\vdots$$

$$y_n = a^{n+1} y_0 + \sum_{k=1}^{n} a^k u_{n-k}$$

Equation:

$$\begin{aligned} E[y_n] &= a^{n+1}E[y_0] + \sum_{k=1}^{n} a^k E[u_{n-k}] \\ &= 0 \end{aligned}$$

Correlation:
Equation:

$$\begin{aligned} E[y_m y_n] &= E\left[\left(a^{m+1}y_0 + \sum_{k=1}^{m} a^k u_{m-k}\right)\left(a^{n+1}y_0 + \sum_{l=1}^{n} a^l u_{n-l}\right)\right] \\ &= E\left[a^{m+n+2}{y_0}^2\right] + E\left[\sum_{k=1}^{m}\sum_{l=1}^{n} a^{k+l}u_{m-k}u_{n-l}\right] \end{aligned}$$

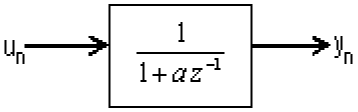
Equation:

$$E[u_{m-k}u_{n-l}] = \begin{cases} {\sigma_n}^2 \text{ if } m-k=n-l \\ 0 \text{ otherwise} \end{cases}$$

If $m > n$, then
Equation:

$$E[y_m y_n] = a^{m+n+2}{\sigma_y}^2 + a^{m-n}{\sigma_u}^2 \sum_{k=1}^{n} a^{2k}$$

If $|a| > 1$, then it's obvious that the process diverges (variance $\rightarrow \infty$). This is equivalent to having a pole outside the unit circle shown in [link].



So, let's assume $|a| < 1$ and hence a stable system. Thus as m and n get large

$$a^{m+n+2}{\sigma_y}^2 \rightarrow 0$$

Now let $m - n = \tau$. Then for m and n large we have
Equation:

$$\begin{aligned} E[y_m y_n] &= a^{\tau} {\sigma_u}^2 \sum_{k=1}^{n} a^{2k} \\ &= \frac{a^{\tau+2} {\sigma_u}^2}{1-a^2} \end{aligned}$$

This shows us how correlated the process is:

$$|a| \rightarrow 1 \Rightarrow \text{heavily correlated (or anticorrelated)}$$

$$|a| \rightarrow 0 \Rightarrow \text{weakly correlated}$$

How can we use this model to help us in our estimation problem?

The Kalman Filter

Let's look at a more general formulation of the problem at hand. Suppose that we have a vector-valued dynamical equation
Equation:

$$y_{n+1} = \boldsymbol{A} y_n + \boldsymbol{b} u_n$$

Where y_n is $p \times 1$ dimensional, $\boldsymbol{A}$ is $p \times p$, and $\boldsymbol{b}$ is $p \times 1$. The initial **state vector** is $Y_0 \sim \mathscr{N}(\mathbf{0}, R_0)$, where R_0 is the covariance matrix and $u_n \sim \mathscr{N}\left(0, {\sigma_u}^2\right)$ iid (white Gaussian **excitation** noise). This reduces to the case we just looked at when $p = 1$. This model could represent a p^{th} order Gauss-Markov process:
Equation:

$$y_{n-1} = a_1 y_n + a_2 y_{n-1} + \ldots + a_p y_{n-p+1} + u_n$$

Define
Equation:

$$yn = \begin{pmatrix} y_{n-p+1} \\ y_{n-p+2} \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}$$

Then,
Equation:

$$\begin{aligned} y_{n+1} &= \boldsymbol{A} y_n + \boldsymbol{b} u_n \\ &= \begin{pmatrix} 0 & 1 & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \dots & 0 & 1 \\ a_1 & a_2 & \dots & \dots & a_{p-1} & a_p \end{pmatrix} \begin{pmatrix} y_{n-p+1} \\ y_{n-p+2} \\ \vdots \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ 1 \end{pmatrix} + u_n \end{aligned}$$

Here $\boldsymbol{A}$ is the **state transition matrix**. Since y_n is a linear combination of Gaussian vectors:
Equation:

$$y_n = \boldsymbol{A}^2 y_0 + \sum_{k=1}^{n} \boldsymbol{A}^{k-1} \boldsymbol{b} u_{n-k}$$

We know that y_n is also Gaussian distributed with mean and covariance $R_n = E\left[y_n {y_n}^T\right], Y_n \sim \mathcal{N}(, R_n)$. The covariance can be recursively computed from the basic state equation:
Equation:

$$R_{n+1} = \boldsymbol{A} R_n \boldsymbol{A}^T + {\sigma_u}^2 \boldsymbol{b} \boldsymbol{b}^T$$

Assume that measurements of the state are available:
Equation:

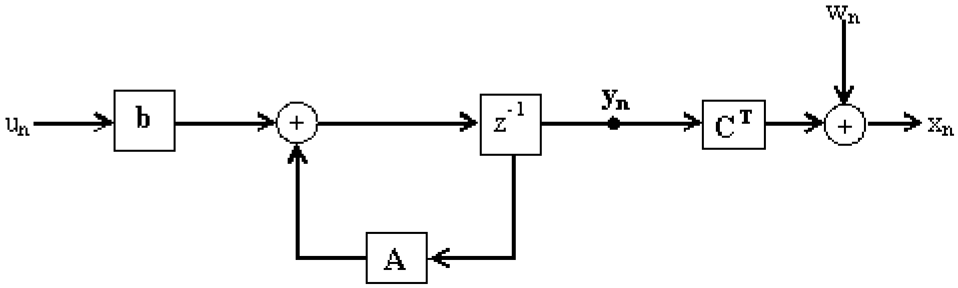
$$\boldsymbol{x}_n = \boldsymbol{C}^T \boldsymbol{y}_n + \boldsymbol{w}_n$$

Where $w_n \sim \mathcal{N}\left(0, {\sigma_w}^2\right)$ iid independant of $\{u_n\}$ (white Gaussian **observation** noise).

For example, if $\boldsymbol{C} = (0 \ldots 01)^T$, then
Equation:

$$\boldsymbol{x}_n = \boldsymbol{y}_n + \boldsymbol{w}_n$$

Where x_n is the observation, y_n is the signal, and w_n is the noise. Since our model for the signal is Gaussian as well as the observation noise, it follows that $x_n \sim \mathcal{N}\left(0, {\sigma_n}^2\right)$, where ${\sigma_n}^2 = \boldsymbol{C}^T R_n \boldsymbol{C} + {\sigma_w}^2$ ([link]).
Block Diagram



Kalman first posed the problem of estimating the state of y_n from the sequence of measurements

$$x_n = \begin{pmatrix} x_0 \\ \vdots \\ x_n \end{pmatrix}$$

To derive the Kalman filter we will call upon the Gauss-Markov Theorem.

First note that the conditional distribution of y_n given x_n is Gaussian:

$$y_n | x_n \sim \mathscr{N}\left(\hat{y}_{n|n}, P_{n|n}\right)$$

Where $\hat{y}_{n|n}$ is the conditional mean and $P_{n|n}$ is the covariance.

We **know** that this is the **form** of the conditional distribution because y_n and x_n are **jointly** Gaussian distributed.

Note:

$$y_n | x_n \sim \mathscr{N}\left(\hat{y}_{n|n}, P_{n|n}\right)$$

where y_n is the signal samples $y_n, \ldots, y_{n-p+1}$, x_n is the observations/measurements $x_n, \ldots, x_{n-p+1}$, and $\hat{y}_{n|n}$ is the best (minimum MSE) estimator of y_n given x_n.

This is all well and good, but we need to know what the conditional mean and covariance are explicitly. So the problem is now to find/compute $\hat{y}_{n|n}$ and $P_{n|n}$. We can take advantage of the recursive state equation to obtain a recursive procedure for this calculation. To begin, consider the "predictor" $\hat{y}_{n|n-1}$:

$$y_n | x_{n-1} \sim \mathscr{N}\left(\hat{y}_{n|n-1}, P_{n|n-1}\right)$$

Where y_n is the signal samples, $\{y_n, \ldots, y_{n-p+1}\}$, x_{n-1} is the observations $\{x_{n-1}, \ldots, x_{n-p}\}$, and $\hat{y}_{n|n-1}$ is the best min MSE estimator of y_n given x_{n-1}. Although we don't know what forms $\hat{y}_{n|n-1}$ and $P_{n|n-1}$ have, we do know two important facts:

1. The predictor $\hat{y}_{n|n-1}$ acts as a sufficient statistic for y_n. That is, we can replace x_{n-1} (the data) with $\hat{y}_{n|n-1}$ (the predictor). In other words, all

the relevant information in x_{n-1} pertaining to y_n is summarized by the predictor $\hat{y}_{n|n-1}$, which is, of course, a function of x_{n-1}.

2. The predictor $\hat{y}_{n|n-1}$ and the prediction error $e_{n|n-1} = y_n - \hat{y}_{n|n-1}$ are orthogonal (the **orthogonality principle** of minimum MSE estimators ⇒ $\left(E\left[\hat{y}_{n|n-1} {e_{n|n-1}}^T\right] = 0\right)$ ⇒ error is orthogonal to estimator).

Moreover,
Equation:

$$\boldsymbol{y_n = \hat{y}_{n|n-1} + e_{n|n-1}}$$

Since all quantities are zero-mean,

$$e_{n|n-1} \sim \mathscr{N}\left(\mathbf{0}, P_{n|n-1}\right)$$

where $P_{n|n-1}$ is the covariance of $y_n | x_{n-1}$ and "variability" of y_n about the predictor $\hat{y}_{n|n-1}$. Therefore,

$$\hat{y}_{n|n-1} \sim \mathscr{N}\left(\mathbf{0}, R_n - P_{n|n-1}\right)$$

Where R_n is the covariance of Y_n. Now suppose that we have the predictor $\hat{y}_{n|n-1}$ computed and a new measurement is made:
Equation:

$$\begin{array}{rcl} \boldsymbol{x_n} & = & \boldsymbol{C^T y_n + w_n} \\ & = & \boldsymbol{C^T \left(\hat{y}_{n|n-1} + e_{n|n-1}\right) + w_n} \end{array}$$

Note: $\hat{y}_{n|n-1}$, $e_{n|n-1}$, and w_n are all **orthogonal**.

We can express all relevant quantities in the matrix equation

Equation:

$$\begin{pmatrix} y_n \\ \hat{y}_{n|n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{I} & 0 \\ \mathbf{0} & \boldsymbol{I} & 0 \\ \boldsymbol{C}^T & \boldsymbol{C}^T & 1 \end{pmatrix} \begin{pmatrix} e_{n|n-1} \\ \hat{y}_{n|n-1} \\ w_n \end{pmatrix}$$

Now because of the orthogonality, the covariance

Equation:

$$E\left[\begin{pmatrix} e_{n|n-1} \\ \hat{y}_{n|n-1} \\ w_n \end{pmatrix} \begin{pmatrix} e_{n|n-1} \\ \hat{y}_{n|n-1} \\ w_n \end{pmatrix}^T\right] = \begin{pmatrix} P_{n|n-1} & \mathbf{0} & 0 \\ \mathbf{0} & R_n - P_{n|n-1} & 0 \\ \mathbf{0} & \mathbf{0} & {\sigma_w}^2 \end{pmatrix}$$

Combining this with the matrix [link] shows that

Equation:

$$E\left[\begin{pmatrix} y_n \\ \hat{y}_{n|n-1} \\ x_n \end{pmatrix} \begin{pmatrix} y_n \\ \hat{y}_{n|n-1} \\ x_n \end{pmatrix}^T\right] = \begin{pmatrix} R_n & \widehat{P}_{n|n-1} & R_n\boldsymbol{C} \\ \widehat{P}_{n|n-1} & \vdots & \dots \\ \boldsymbol{C}^T R_n & \vdots & S_n \end{pmatrix}$$

Note: $(y_n \hat{y}_{n|n-1} x_n)^T$ are jointly Gaussian with the covariance in [link] and means zero.

Where

Equation:

$$\widehat{P}_{n|n-1} = R_n - \widehat{P}_{n|n-1}$$

Equation:

$$S_n = \begin{pmatrix} \widehat{P}_{n|n-1} & \widehat{P}_{n|n-1}\boldsymbol{C} \\ \boldsymbol{C}^T\widehat{P}_{n|n-1} & {\sigma_w}^2 \end{pmatrix}$$

We now have all the quantities necessary to compute our recursive estimator using the Gauss-Markov Theorem.

We will now derive a recursion for conditional distribution of y_n given $\hat{y}_{n|n-1}$ (best estimate based on past observations) and x_n (current observation). We know that $y_n\Big|\left(\hat{y}_{n|n-1}, x_n\right)$ is Gaussian (since all quantities are jointly Gaussian). Let's denote this conditional distribution by

$$y_n\Big|\left(\hat{y}_{n|n-1}, x_n\right) \sim \mathscr{N}\left(\hat{y}_{n|n}, P_{n|n}\right)$$

Applying the Gauss-Markov Theorem we find
Equation:

$$\hat{y}_{n|n} = \begin{pmatrix} \widehat{P}_{n|n-1} \\ R_n\boldsymbol{C} \end{pmatrix}^T {S_n}^{-1} \begin{pmatrix} \hat{y}_{n|n-1} \\ x_n \end{pmatrix}$$

which is the best estimator of y_n given $\hat{y}_{n|n-1}$ and x_n. The inverse of S_n is given by
Equation:

$${S_n}^{-1} = \begin{pmatrix} \left(\widehat{P}_{n|n-1}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + {\gamma_n}^{-1}\begin{pmatrix} -\boldsymbol{C} \\ 1 \end{pmatrix}\begin{pmatrix} -\boldsymbol{C}^T \\ 1 \end{pmatrix}^T$$

where
Equation:

$$\begin{aligned} {\gamma_n}^{-1} &= {\sigma_n}^2 - \boldsymbol{C}^T \left(R_n - P_{n|n-1}\right) \boldsymbol{C} \\ &= \boldsymbol{C}^T P_{n|n-1} \boldsymbol{C} + {\sigma_w}^2 \end{aligned}$$

Note: ${\sigma_n}^2 = \boldsymbol{C}^T R_n \boldsymbol{C} + {\sigma_w}^2$

Substituting this inverse formula into [link] yields
Equation:

$$\hat{y}_{n|n} = \hat{y}_{n|n-1} + P_{n|n-1} \boldsymbol{C} {\gamma_n}^{-1} \left(\boldsymbol{x}_n - \boldsymbol{C}^T \hat{y}_{n|n-1}\right)$$

The Gauss-Markov Theorem also gives us an expression for $P_{n|n}$:
Equation:

$$P_{n|n} = R_n - \begin{pmatrix} \widehat{P}_{n|n-1} \\ R_n \boldsymbol{C} \end{pmatrix}^T {S_n}^{-1} \begin{pmatrix} \widehat{P}_{n|n-1} \\ \boldsymbol{C}^T R_n \end{pmatrix}$$

and upon substituting [link] for ${S_n}^{-1}$ we get
Equation:

$$P_{n|n} = P_{n|n-1} - {\gamma_n}^{-1} P_{n|n-1} \boldsymbol{C} \boldsymbol{C}^T P_{n|n-1}$$

Note that both expressions contain the quantity
Equation:

$$K_n = P_{n|n-1} \boldsymbol{C} {\gamma_n}^{-1}$$

which is the so-called **Kalman gain**.

Using the Kalman gain, the **Kalman recursions** are given by
Equation:

$$\boldsymbol{\hat{y}}_{n|n} = \boldsymbol{\hat{y}}_{n|n-1} + \boldsymbol{K}_n \left(\boldsymbol{x}_n - \boldsymbol{C}^T \boldsymbol{\hat{y}}_{n|n-1}\right)$$

Equation:

$$P_{n|n} = P_{n|n-1} - \gamma_n \boldsymbol{K}_n \boldsymbol{K}_n{}^T$$

The recursions are complete except for definitions of $\hat{y}_{n|n-1}$ and $P_{n|n-1}$.
Equation:

$$\begin{aligned} \boldsymbol{\hat{y}}_{n|n-1} &= E[\boldsymbol{y}_n | \boldsymbol{x}_{n-1}] \\ &= E[\boldsymbol{A}\boldsymbol{y}_{n-1} + \boldsymbol{b}u_{n-1} | x_{n-1}] \\ &= \boldsymbol{A}\boldsymbol{\hat{y}}_{n-1|n-1} \end{aligned}$$

Equation:

$$\begin{aligned} P_{n|n-1} &= E\left[\left(y_n - \hat{y}_{n|n-1}\right)\left(y_n - \hat{y}_{n|n-1}\right)^T\right] \\ &= \boldsymbol{A}P_{n-1|n-1}\boldsymbol{A}^T + \sigma_n{}^2 \boldsymbol{b}\boldsymbol{b}^T \end{aligned}$$

Now we can summarize the **Kalman filter**:

1. [link], where $\hat{y}_{n|n}$ is the best estimate if y_n given observations up to time n.
2. [link]
3. [link]
4. [link]
5. [link]
6. [link]

Measurements/observation model:

$$x_n = y_n + w_n$$

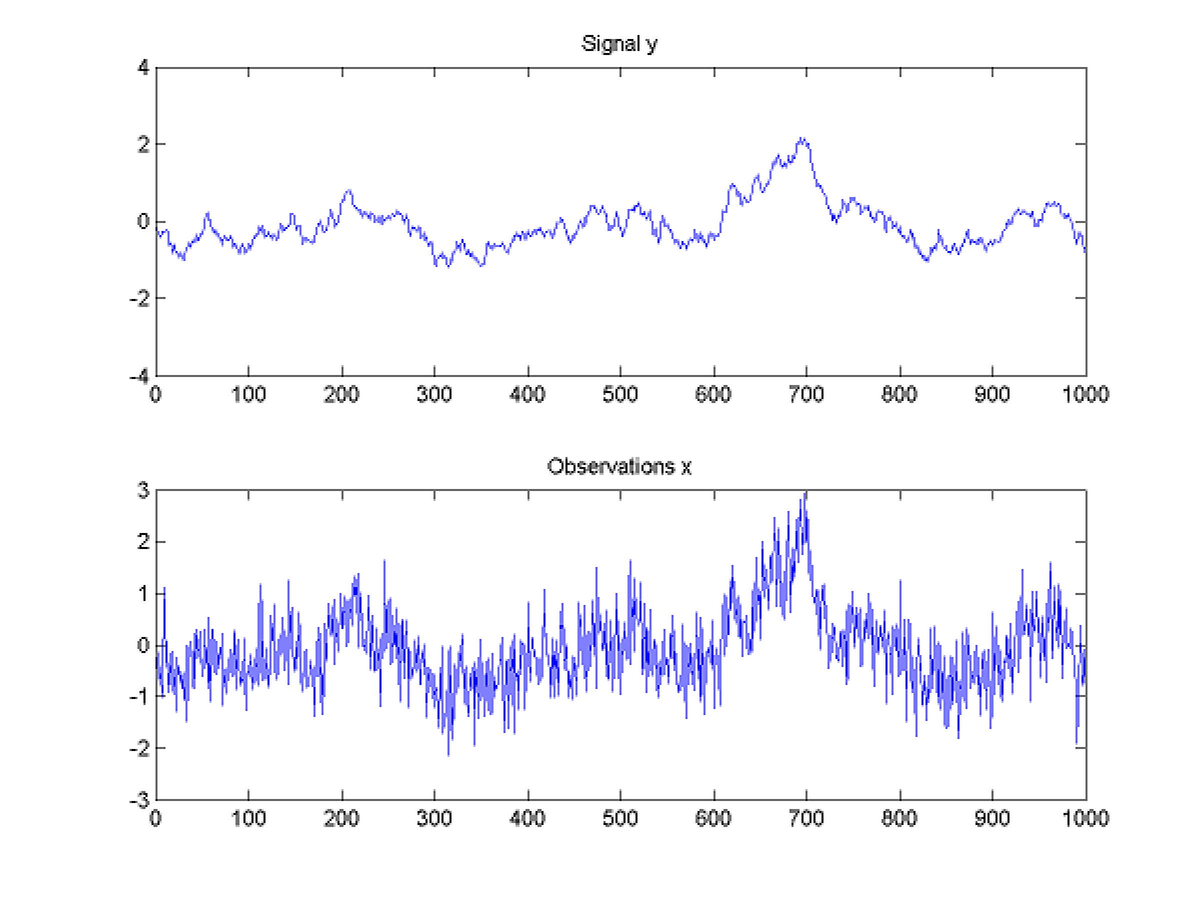
$$w_n \sim \mathcal{N}\left(0, {\sigma_w}^2\right)$$

($\boldsymbol{C} = 1$).

Example:
First-order Gauss-Markov Process

$$y_{n+1} = a y_n + u_n$$

Where y_{n+1} is a time-varying voltage, $u_n \sim \mathcal{N}\left(0, {\sigma_n}^2\right)$, and $\sigma_u = 0.1$. ($a = 0.99$) $\Rightarrow$ highly correlated process. ($\boldsymbol{A} = a$, $\boldsymbol{b} = 1$).

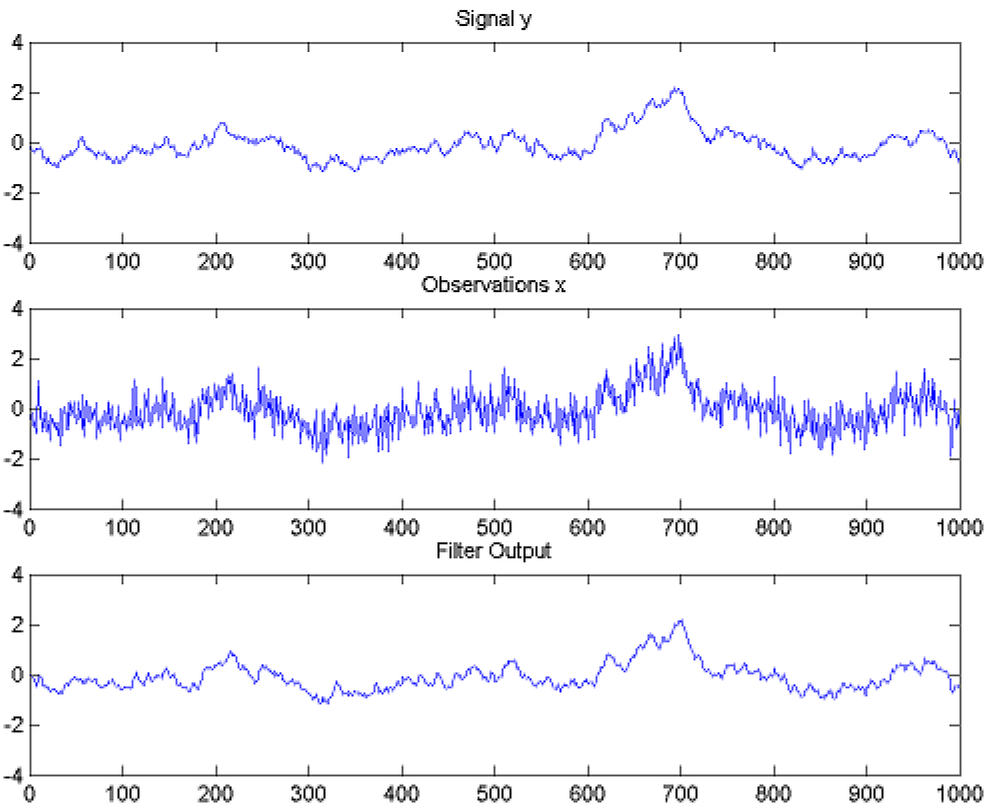


Kalman Filtering Equations

1. $P_{n|n-1} = a^2 P_{n-1|n-1} + {\sigma_n}^2$ ($q(n)$ in MATLAB code)
2. ${\gamma_n}^{-1} = P_{n|n-1} + {\sigma_w}^2$ ($g(n)$ in MATLAB code)

3. $P_{n|n} = P_{n|n-1} - {\gamma_n}^{-1} {P_{n|n-1}}^2$ ($p(n)$ in MATLAB code)
4. $K_n = P_{n|n-1} {\gamma_n}^{-1}$ ($k(n)$ in MATLAB code)
5. $\hat{y}_{n|n-1} = a\hat{y}_{n-1|n-1}$ (py (n) in MATLAB code)
6. $\hat{y}_{n|n} = \hat{y}_{n|n-1} + K_n \left(x_n - \hat{y}_{n|n-1}\right)$ (ey (n) in MATLAB code)

Initialization: ey $(1) = 0$, $q(1) = {\sigma_u}^2$



Introduction to Adaptive Filtering

The Kalman filter is just one of many **adaptive** filtering (or estimation) algorithms. Despite its elegant derivation and often excellent performance, the Kalman filter has two drawbacks:

1. The derivation and hence performance of the Kalman filter depends on the accuracy of the a priori assumptions. The performance can be less than impressive if the assumptions are erroneous.
2. The Kalman filter is fairly computationally demanding, requiring $O\left(P^2\right)$ operations per sample. This can limit the utility of Kalman filters in high rate real time applications.

As a popular alternative to the Kalman filter, we will investigate the so-called **least-mean-square** (LMS) adaptive filtering algorithm.

The principle advantages of LMS are

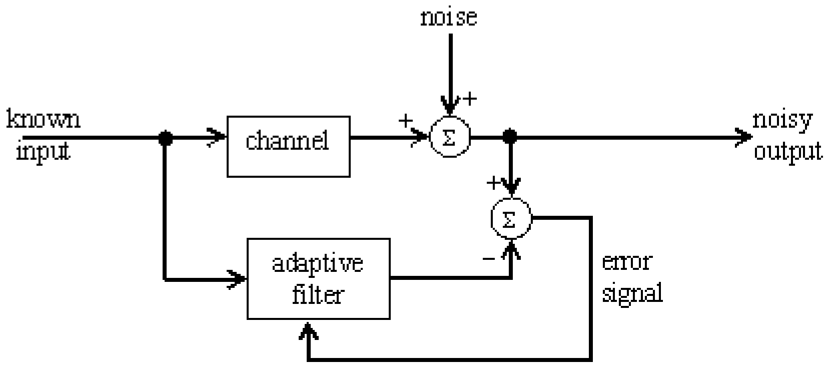
1. No prior assumptions are made regarding the signal to be estimated.
2. Computationally, LMS is very efficient, requiring $O(P)$ per sample.

The price we pay with LMS instead of a Kalman filter is that the rate of convergence and adaptation to sudden changes is slower for LMS than for the Kalman filter (with correct prior assumptions).

Adaptive Filtering Applications

Channel/System Identification

Channel/System Identification

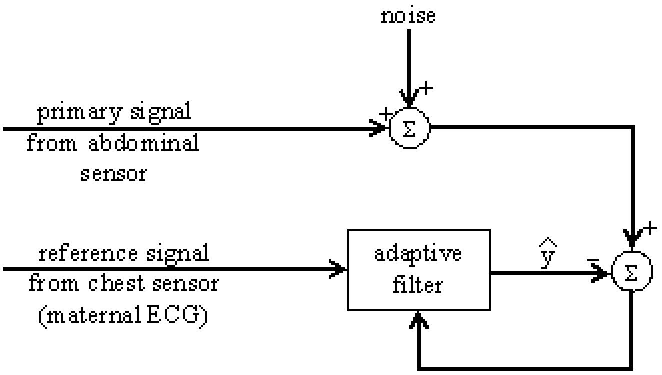


Noise Cancellation

Suppression of maternal ECG component in fetal ECG ([link]).

[missing_resource: .png]

Cancelling maternal heartbeat in fetal electrocardiography (ECG): position of leads.



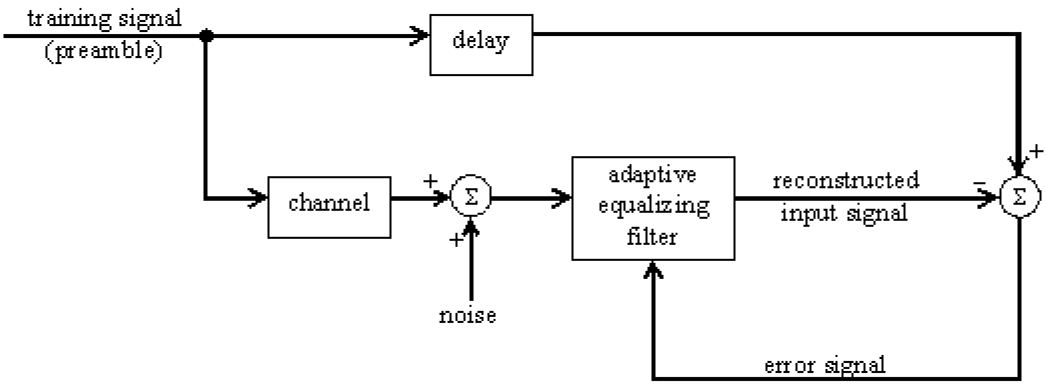
$\hat{y}$ is an estimate of the maternal ECG signal present in the abdominal signal ([link]).

[missing_resource: .png]

Results of fetal ECG experiment (bandwidth, 3-35Hz; sampling rate, 256Hz): (a)reference input (chest lead); (b)primary input (abdominal lead); (c)noise-canceller output.

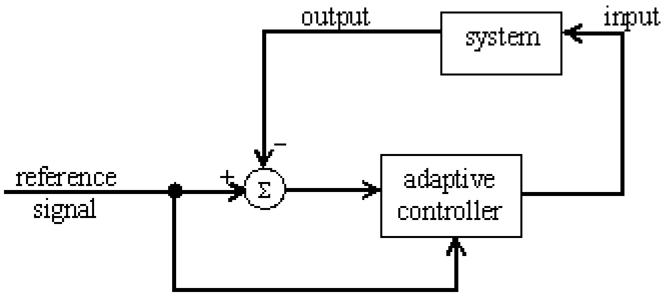
Channel Equalization

Channel Equalization



Adaptive Controller

Adaptive Controller



Here, the reference signal is the desired output. The adaptive controller adjusts the controller gains (filter weights) to keep them appropriate to the system as it changes over time.

Iterative Minimization

Most adaptive filtering alogrithms (LMS included) are modifications of standard iterative procedures for solving minimization problems in a **real-time** or **on-line** fashion. Therefore, before deriving the LMS algorithm we will look at iterative methods of minimizing error criteria such as MSE.

Conider the following set-up:

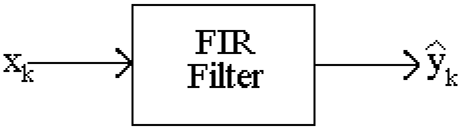
$$x_k \text{ : observation}$$

$$y_k \text{ : signal to be estimated}$$

Linear estimator

Equation:

$$\hat{y}_k = w_1 x_k + w_2 x_{k-1} + \ldots + w_p x_{k-p+1}$$



Impulse response of the filter:

$$\ldots, 0, 0, w_1, w_2, \ldots w_p, 0, 0, \ldots$$

Vector notation

Equation:

$$\hat{y}_k = x_k{}^T \boldsymbol{w}$$

Where

$$x_k = \begin{pmatrix} x_k \\ x_{k-1} \\ \vdots \\ x_{k-p+1} \end{pmatrix}$$

and

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix}$$

Error signal

Equation:

$$\begin{aligned} e_k &= y_k - \hat{y}_k \\ &= y_k - {x_k}^T \boldsymbol{w} \end{aligned}$$

Assumptions

(x_k, y_k) are jointly stationary with zero-mean.

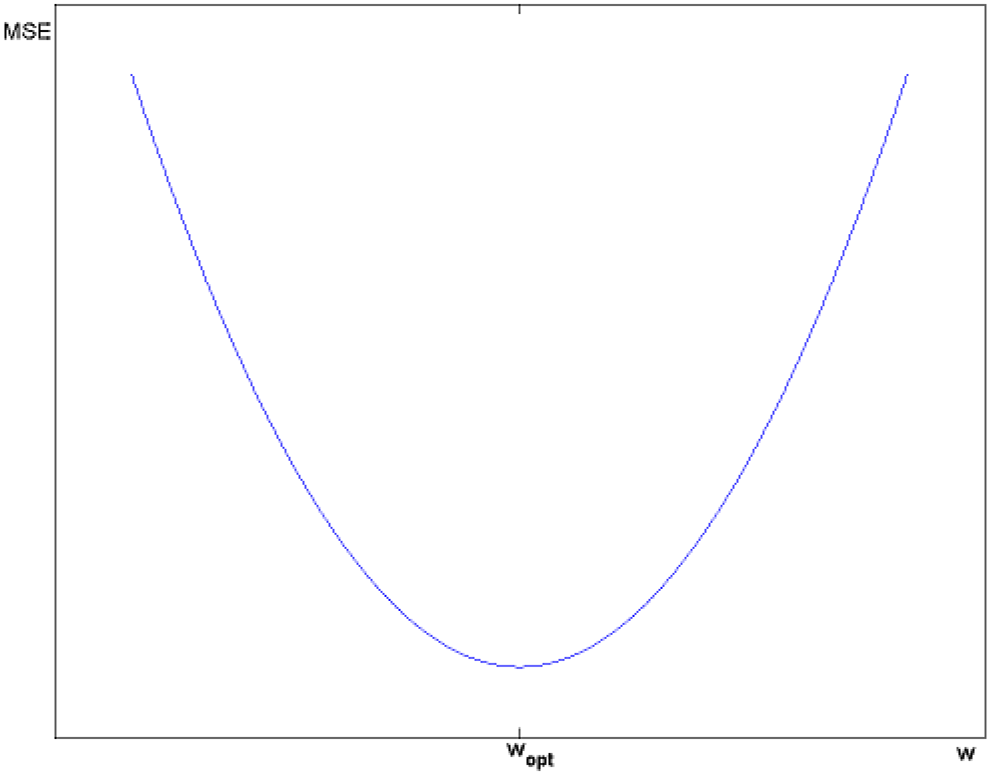
MSE

Equation:

$$\begin{aligned}
E\left[{e_k}^2\right] &= E\left[\left(y_k - {x_k}^T \boldsymbol{w}\right)^2\right] \\
&= E\left[{y_k}^2\right] - 2\boldsymbol{w}^T E[x_k y_k] + \boldsymbol{w}^T E\left[x_k {x_k}^T\right] \boldsymbol{w} \\
&= R_{\mathrm{yy}} - 2\boldsymbol{w}^T R_{\mathrm{xy}} + \boldsymbol{w}^T R_{\mathrm{xx}} \boldsymbol{w}
\end{aligned}$$

Where R_{yy} is the variance of ${y_k}^2$, R_{xx} is the covariance matrix of x_k, and $R_{\mathrm{xy}} = E[x_k y_k]$ is the cross-covariance between x_k and y_k

Note: The MSE is quadratic in $\boldsymbol{W}$ which implies the MSE surface is "bowl" shaped with a unique minimum point ([link]).



Optimum Filter

Minimize MSE:
Equation:

$$\left(\frac{\partial E\left[{e_{k}}^{2}\right]}{\partial \boldsymbol{w}}=2R_{\mathrm{xy}}+2R_{\mathrm{xx}}\boldsymbol{w}=0\right)\Rightarrow\left(w_{\mathrm{opt}}={R_{\mathrm{xx}}}^{-1}R_{\mathrm{xy}}\right)$$

Notice that we can re-write [link] as
Equation:

$$E\left[\boldsymbol{x}_{k}{\boldsymbol{x}_{k}}^{T}\boldsymbol{w}\right]=E[\boldsymbol{x}_{k}\boldsymbol{y}_{k}]$$

or
Equation:

$$\begin{array}{rcl} E\left[\boldsymbol{x}_{k}\left(\boldsymbol{y}_{k}-{\boldsymbol{x}_{k}}^{T}\boldsymbol{w}\right)\right] & = & E[\boldsymbol{x}_{k}\boldsymbol{e}_{k}] \\ & = & 0 \end{array}$$

Which shows that the error signal is orthogonal to the input x_k (by the **orthogonality principle** of minimum MSE estimator).

Steepest Descent

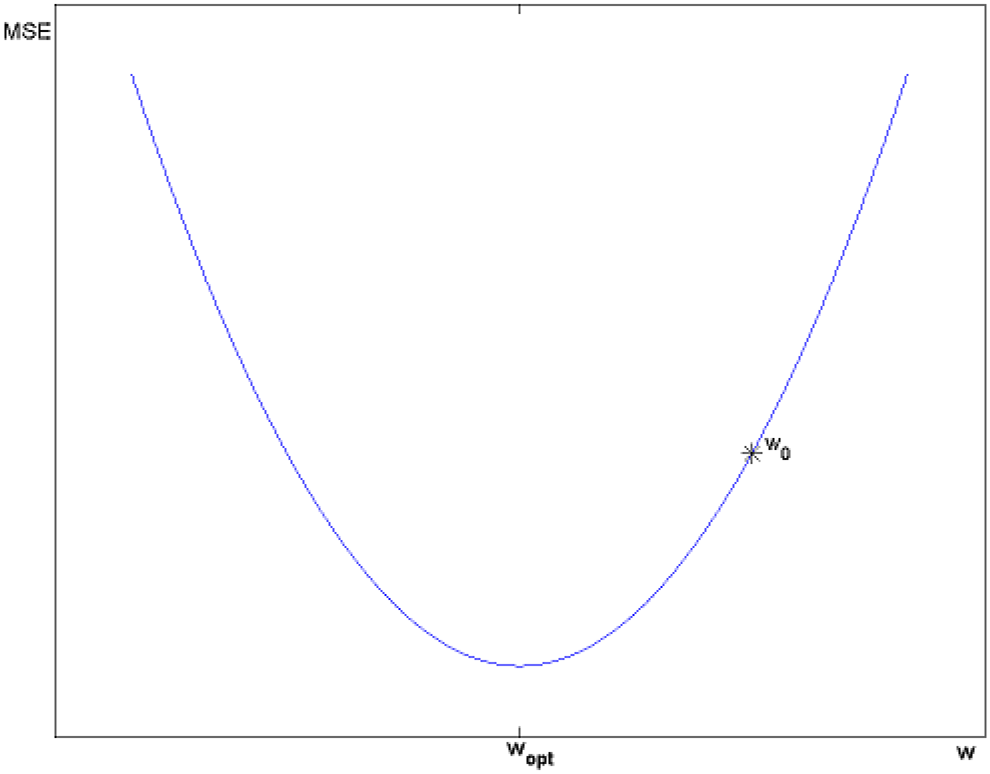
Although we can easily determine w_{opt} by solving the system of equations
Equation:

$$R_{\mathrm{xx}}\boldsymbol{w}=R_{\mathrm{xy}}$$

Let's look at an iterative procedure for solving this problem. This will set the stage for our adaptive filtering algorithm.

We want to minimize the MSE. The idea is simple. Starting at some initial weight vector w_0, iteratively adjust the values to decrease the MSE ([link]).

In One-Dimension

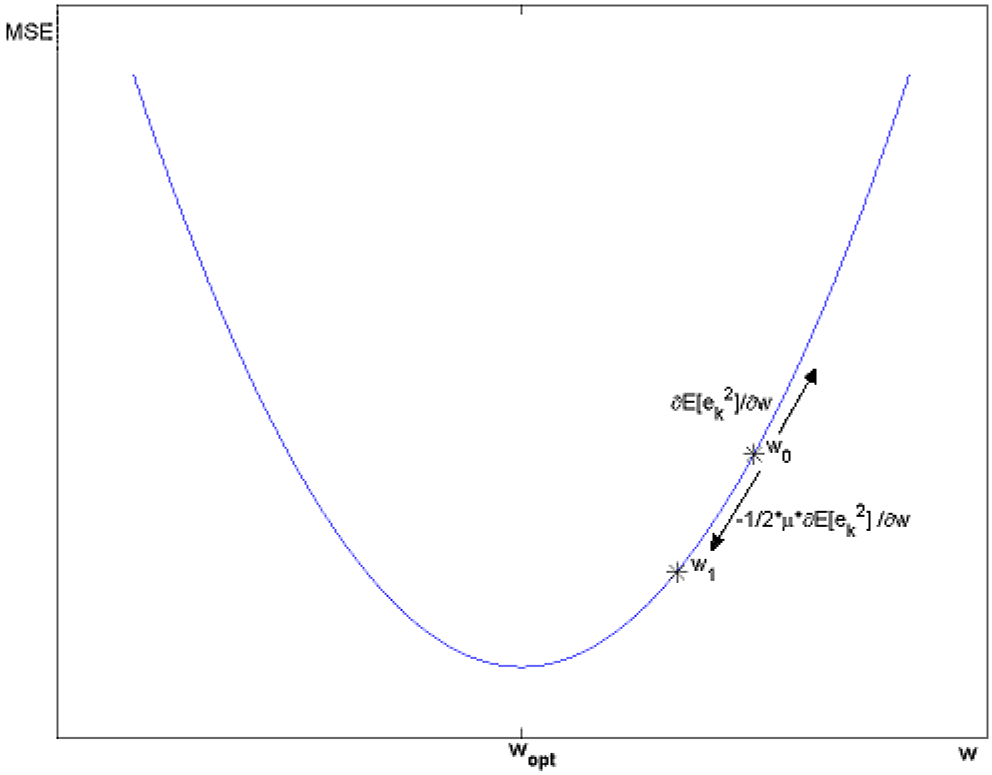


We want to **move** w_0 towards the optimal vector w_{opt}. In order to move in the correct direction, we must move **downhill** or in the direction opposite to the gradient of the MSE surface at the point w_0. Thus, a natural and simple adjustment takes the form

Equation:

$$w_1 = w_0 - \frac{1}{2}\mu \frac{\partial E\left[{e_k}^2\right]}{\partial \boldsymbol{w}}\Bigg|_{\boldsymbol{w}=w_0}$$

Where μ is the step size and tells us how far to move in the negative gradient direction ([link]).



Generalizing this idea to an iterative strategy, we get
Equation:

$$w_k = w_{k-1} - \frac{1}{2}\mu \frac{\partial E\left[{e_k}^2\right]}{\partial \boldsymbol{w}}\bigg|_{\boldsymbol{w}=w_{k-1}}$$

and we can repeatedly update $\boldsymbol{w}$: $w_0, w_1, \ldots, w_k$. Hopefully each subsequent w_k is closer to w_{opt}. Does the procedure converge? Can we adapt it to an on-line, real-time, dynamic situation in which the signals may not be stationary?

LMS Algorithm Analysis

Objective

Minimize instantaneous squared error
Equation:

$${e_k}^2(\boldsymbol{w}) = \left(y_k - {x_k}^T\boldsymbol{w}\right)^2$$

LMS Algorithm

Equation:

$$\widehat{w}_k = \widehat{w}_{k-1} + \boldsymbol{\mu x_k e_k}$$

Where w_k is the new weight vector, w_{k-1} is the old weight vector, and $\boldsymbol{\mu x_k e_k}$ is a small step in the instantaneous error gradient direction.

Interpretation in Terms of Weight Error Vector

Define
Equation:

$$v_k = w_k - w_{\text{opt}}$$

Where w_{opt} is the optimal weight vector and
Equation:

$$\varepsilon_k = y_k - {x_k}^T w_{\text{opt}}$$

where ε_k is the minimum error. The stochastic difference equation is:
Equation:

$$v_k = \boldsymbol{I}v_{k-1} + \mu x_k \varepsilon_k$$

Convergence/Stability Analysis

Show that (tightness)
Equation:

$$\underset{B \to \infty}{\text{limit}} \ \max\{\Pr[\| v_k \| \geq B]\} = 0$$

With probability 1, the weight error vector is bounded for all k.

Chebyshev's inequality is
Equation:

$$\Pr[\| v_k \| \geq B] \leq \frac{E\left[(\| v_k \|)^2\right]}{B^2}$$

and
Equation:

$$\Pr[\| v_k \| \geq B] = \frac{1}{B^2}\left((\| E[v_k] \|)^2 + \sigma(v_k)^2\right)$$

where $(\| E[v_k] \|)^2$ is the squared bias. If $(\| E[v_k] \|)^2 + \sigma(v_k)^2$ is finite for all k, then $\underset{B \to \infty}{\text{limit}} \ \Pr[\| v_k \| \geq B] = 0$ for all k.

Also,
Equation:

$$\sigma(v_k)^2 = \text{tr}\left(E\left[v_k {v_k}^T\right]\right)$$

Therefore $\sigma(v_k)^2$ is finite if the diagonal elements of $\mathit{\Gamma}_k \equiv E\left[v_k {v_k}^T\right]$ are bounded.

Convergence in Mean

$\| \; E[v_k] \; \| \to 0$ as $k \to \infty$. Take expectation of [link] using smoothing property to simplify the calculation. We have convergence in mean if

1. R_{xx} is positive definite (invertible).
2. $\mu < \frac{2}{\lambda_{\max}(R_{\mathrm{xx}})}$.

Bounded Variance

Show that $\mathit{\Gamma}_k = E\left[v_k {v_k}^T\right]$, the weight vector error covariance is bounded for all k.

Note: We could have $E[v_k] \to 0$, but $\sigma(v_k)^2 \to \infty$; in which case the algorithm would not be stable.

Recall that it is fairly straightforward to show that the diagonal elements of the transformed covariance $C_k = \boldsymbol{U} \mathit{\Gamma}_k \boldsymbol{U}^T$ tend to zero if $\mu < \frac{1}{\lambda_{\max}(R_{\mathrm{xx}})}$ ($\boldsymbol{U}$ is the eigenvector matrix of R_{xx}; $R_{\mathrm{xx}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$). The diagonal elements of C_k were denoted by $\gamma_{k,i} \forall i, i = \{1, \ldots, p\} : (i = \{1, \ldots, p\})$.

Note: $\sigma(v_k)^2 = \mathrm{tr}\left(\mathit{\Gamma}_k\right) = \mathrm{tr}\left(\boldsymbol{U}^T C_k \boldsymbol{U}\right) = \mathrm{tr}\left(C_k\right) = \sum_{i=1}^{p} \gamma_{k,i}$

Thus, to guarantee boundedness of $\sigma(v_k)^2$ we need to show that the "steady-state" values $\gamma_{k,i} \to (\gamma_i < \infty)$.

We showed that

Equation:

$$\gamma_i = \frac{\mu\left(\alpha + {\sigma_\varepsilon}^2\right)}{2 \times (1 - \mu\lambda_i)}$$

where ${\sigma_\varepsilon}^2 = E\left[{\varepsilon_k}^2\right]$, λ_i is the i^{th} eigenvalue of R_{xx} ($R_{\text{xx}} = \boldsymbol{U} \begin{matrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_p \end{matrix} \boldsymbol{U}^T$), and $\alpha = \frac{c{\sigma_\varepsilon}^2}{1-c}$.

Equation:

$$0 < c = \frac{1}{2}\sum_{i=1}^{p} \frac{\mu\lambda_i}{1 - \mu\lambda_i} < 1$$

We found a sufficient condition for μ that guaranteed that the steady-state γ_i's (and hence σ$(v_k)^2$) are bounded:

$$\mu < \frac{\frac{2}{3}}{\sum_{i=1}^{p} \lambda_i}$$

Where $\sum_{i=1}^{p} \lambda_i = \text{tr}\left(R_{\text{xx}}\right)$ is the input vector energy.

With this choice of μ we have:

1. convergence in mean
2. bounded steady-state variance

This implies

Equation:

$$\underset{B \to \infty}{\text{limit}} \ \max\left\{\Pr[\| \ v_k \ \| \geq B]\right\} = 0$$

In other words, the LMS algorithm is stable about the optimum weight vector w_{opt}.

Learning Curve

Recall that

Equation:

$$e_k = y_k - {x_k}^T w_{k-1}$$

and [link]. These imply

Equation:

$$e_k = \varepsilon_k - {x_k}^T v_{k-1}$$

where $v_k = w_k - w_{\mathrm{opt}}$. So the MSE

Equation:

$$\begin{aligned} E\left[{e_k}^2\right] &= {\sigma_\varepsilon}^2 + E\left[{v_{k-1}}^T x_k {x_k}^T v_{k-1}\right] \\ &= {\sigma_\varepsilon}^2 + E\left[E\left[{v_{k-1}}^T x_k {x_k}^T v_{k-1} \,|\, x_n \varepsilon_n \forall n, n < k : (n < k)\right]\right] \\ &= {\sigma_\varepsilon}^2 + E\left[{v_{k-1}}^T R_{\mathrm{xx}} v_{k-1}\right] \\ &= {\sigma_\varepsilon}^2 + E\left[\mathrm{tr}\left(R_{\mathrm{xx}} v_{k-1} {v_{k-1}}^T\right)\right] \\ &= {\sigma_\varepsilon}^2 + \mathrm{tr}\left(R_{\mathrm{xx}} \Gamma_{k-1}\right) \end{aligned}$$

Where $(\mathrm{tr}\left(R_{\mathrm{xx}} \Gamma_{k-1}\right) \equiv \alpha_{k-1}) \rightarrow \left(\alpha = \frac{c{\sigma_\varepsilon}^2}{1-c}\right)$. So the limiting MSE is

Equation:

$$\begin{aligned} \varepsilon_\infty &= \lim_{k \to \infty} E\left[{e_k}^2\right] \\ &= {\sigma_\varepsilon}^2 + \frac{c{\sigma_\varepsilon}^2}{1-c} \\ &= \frac{{\sigma_\varepsilon}^2}{1-c} \end{aligned}$$

Since $0 < c < 1$ was required for convergence, $\varepsilon_\infty > {\sigma_\varepsilon}^2$ so that we see noisy adaptation leads to an MSE larger than the optimal
Equation:

$$\begin{aligned} E\left[{\varepsilon_k}^2\right] &= E\left[\left(y_k - {x_k}^T w_{\mathrm{opt}}\right)^2\right] \\ &= {\sigma_\varepsilon}^2 \end{aligned}$$

To quantify the increase in the MSE, define the so-called **misadjustment**:
Equation:

$$\begin{aligned} M &= \frac{\varepsilon_\infty - {\sigma_\varepsilon}^2}{{\sigma_\varepsilon}^2} \\ &= \frac{\varepsilon_\infty}{{\sigma_\varepsilon}^2} - 1 \\ &= \frac{\alpha}{{\sigma_\varepsilon}^2} \\ &= \frac{c}{1-c} \end{aligned}$$

We would of course like to keep M as small as possible.

Learning Speed and Misadjustment Trade-off

Fast adaptation and quick convergence require that we take steps as large as possible. In other words, learning speed is proportional to μ; larger μ means faster convergence. How does μ affect the misadjustment?

To guarantee convergence/stability we require

$$\mu < \frac{\frac{2}{3}}{\sum_{i=1}^{p} \lambda_i(R_{\mathrm{xx}})}$$

Let's assume that in fact $\mu \ll \frac{1}{\sum_{i=1}^{p} \lambda_i}$ so that there is no problem with convergence. This condition implies $\mu \ll \frac{1}{\lambda_i}$ or $\mu\lambda_i \ll 1 \forall i, i = \{1, \ldots, p\} : (i = \{1, \ldots, p\})$. From here we see that

Equation:

$$c = \frac{1}{2}\sum_{i=1}^{p}\frac{\mu\lambda_i}{1-\mu\lambda_i} \simeq \frac{1}{2}\mu\sum_{i=1}^{p}\lambda_i \ll 1$$

This misadjustment
Equation:

$$M = \frac{c}{1-c} \simeq c = \frac{1}{2}\mu\sum_{i=1}^{p}\lambda_i$$

This shows that larger step size μ leads to larger misadjustment.

Since we still have convergence in mean, this essentially means that with a larger step size we "converge" faster but have a larger variance (rattling) about w_{opt}.

Summary

small μ implies

- small misadjustment in steady-state
- slow adaptation/tracking

large μ implies

- large misadjustment in steady-state
- fast adaptation/tracking

Example:

$$w_{\text{opt}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$x_k \sim \mathscr{N}\left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

$$y_k = {x_k}^T w_{\text{opt}} + \varepsilon_k$$

$$\varepsilon_k \sim \mathscr{N}(0, 0.01)$$

LMS Algorithm

initialization $w_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and
$w_k = w_{k-1} + \mu x_k e_k \forall k, k \geq 1 : (k \geq 1)$, where
$e_k = y_k - {x_k}^T w_{k-1}$

Learning Curve

[missing_resource: .png]

$$\mu = 0.05$$

LMS Learning Curve

[missing_resource: .png]

$$\mu = 0.3$$

Comparison of Learning Curves

[missing_resource: .png]